sensors

Special Issue Reprint

Multimodal Perception Modeling Based on Advanced Computational Technologies

Edited by
Manuel Gil-Martín, Rubén San-Segundo and Fernando Fernández-Martínez

mdpi.com/journal/sensors

MDPI

Multimodal Perception Modeling Based on Advanced Computational Technologies

Multimodal Perception Modeling Based on Advanced Computational Technologies

Guest Editors

Manuel Gil-Martín
Rubén San-Segundo
Fernando Fernández-Martínez

MDPI

Basel • Beijing • Wuhan • Barcelona • Belgrade • Novi Sad • Cluj • Manchester

Guest Editors

Manuel Gil-Martín
Department of Electronics Engineering
Universidad Politécnica de Madrid
Madrid
Spain

Rubén San-Segundo
Department of Electronics Engineering
Universidad Politécnica de Madrid
Madrid
Spain

Fernando Fernández-Martínez
Department of Electronics Engineering
Universidad Politécnica de Madrid
Madrid
Spain

Editorial Office

MDPI AG
Grosspeteranlage 5
4052 Basel, Switzerland

This is a reprint of the Special Issue, published open access by the journal *Sensors* (ISSN 1424-8220), freely accessible at: https://www.mdpi.com/journal/sensors/special_issues/0871M3OOZV.

For citation purposes, cite each article independently as indicated on the article page online and as indicated below:

Lastname, A.A.; Lastname, B.B. Article Title. *Journal Name* **Year**, *Volume Number*, Page Range.

ISBN 978-3-7258-8547-3 (Hbk)
ISBN 978-3-7258-8548-0 (PDF)
https://doi.org/10.3390/books978-3-7258-8548-0

© 2026 by the authors. Articles in this reprint are Open Access and distributed under the Creative Commons Attribution (CC BY) license. The reprint as a whole is distributed by MDPI under the terms and conditions of the Creative Commons Attribution-NonCommercial-NoDerivs (CC BY-NC-ND) license (https://creativecommons.org/licenses/by-nc-nd/4.0/).

Contents

Chenjia Huang, Yanli Chen, Bocheng Zhou, Xiuqi Cai, Ziying Zhai, Jiarui Zhang, et al.
LightEdu-Net: Noise-Resilient Multimodal Edge Intelligence for Student-State Monitoring in Resource-Limited Environments
Reprinted from: *Sensors* **2025**, *25*, 7529, https://doi.org/10.3390/s25247529 . . . . . **1**

Vanessa Camilleri, Reno Yuri Camilleri, Mark Fialovszky, Daniel Pace, Dylan Seychell and Matthew Montebello
Towards Explainable Multimodal Sensing for Swimming Analysis: Early Findings from the SWIM-360 Project
Reprinted from: *Sensors* **2025**, *25*, 7047, https://doi.org/10.3390/s25227047 . . . . . **23**

Siyu Liang, Yunan Li, Wentian Xin, Huizhou Chen, Xujie Liu, Kang Liu, et al.
Generative Sign-Description Prompts with Multi-Positive Contrastive Learning for Sign Language Recognition
Reprinted from: *Sensors* **2025**, *25*, 5957, https://doi.org/10.3390/s25195957 . . . . . **53**

Yanlin Chen, Yingqiu Yang, Zeyu Lan, Xinyuan Chen, Haoyuan Zhan, Lingxi Yu, et al.
Image Sensor-Supported Multimodal Attention Modeling for Educational Intelligence
Reprinted from: *Sensors* **2025**, *25*, 5640, https://doi.org/10.3390/s25185640 . . . . . **75**

Jinkwon Hwang, Mingyu Jeon, Junyeong Kim
Multimodal Latent Representation Learning for Video Moment Retrieval
Reprinted from: *Sensors* **2025**, *25*, 4528, https://doi.org/10.3390/s25144528 . . . . . **94**

Gabriela Hilario-Acuapan, Keny Ordaz-Hernández, Mario Castelán, Ismael Lopez-Juarez
Toward a Recognition System for Mexican Sign Language: Arm Movement Detection
Reprinted from: *Sensors* **2025**, *25*, 3636, https://doi.org/10.3390/s25123636 . . . . . **109**

Yanlin Chen, Chenjia Huang, Shumiao Gao, Yifan Lyu, Xinyuan Chen, Shen Liu, et al.
A Multimodal Deep Learning Approach for Legal English Learning in Intelligent Educational Systems
Reprinted from: *Sensors* **2025**, *25*, 3397, https://doi.org/10.3390/s25113397 . . . . . **135**

Toshinori Oba, Midori Tanaka and Takahiko Horiuchi
Analysis of Pleasure and Displeasure in Harmony Between Colored Light and Fragrance by the Left and Right OFC Response Differences
Reprinted from: *Sensors* **2025**, *25*, 2230, https://doi.org/10.3390/s25072230 . . . . . **165**

Iván Martín-Fernández, Sergio Esteban-Romero, Fernando Fernández-Martínez and Manuel Gil-Martín
Parameter-Efficient Adaptation of Large Vision—Language Models for Video Memorability Prediction
Reprinted from: *Sensors* **2025**, *25*, 1661, https://doi.org/10.3390/s25061661 . . . . . **185**

Abdulrahman Alkurdi, Maxine He, Jonathan Cerna, Jean Clore, Richard Sowers, Elizabeth T. Hsiao-Wecksler, et al.
Extending Anxiety Detection from Multimodal Wearables in Controlled Conditions to Real-World Environments
Reprinted from: *Sensors* **2025**, *25*, 1241, https://doi.org/10.3390/s25041241 . . . . . **201**

sensors

MDPI

Article

LightEdu-Net: Noise-Resilient Multimodal Edge Intelligence for Student-State Monitoring in Resource-Limited Environments

Chenjia Huang [1], Yanli Chen [1,2], Bocheng Zhou [1], Xiuqi Cai [1], Ziying Zhai [2], Jiarui Zhang [2] and Yan Zhan [1,3,*]

[1] National School of Development, Peking University, Beijing 100871, China
[2] China Agricultural University, Beijing 100083, China
[3] Artificial Intelligence Research Institute, Tsinghua University, Beijing 100084, China
* Correspondence: yan_zhan_thu@163.com

Abstract

Multimodal perception for student-state monitoring is difficult to deploy in rural classrooms because sensors are noisy and computing resources are highly constrained. This work targets these challenges by enabling noise-resilient, multimodal, real-time student-state recognition on low-cost edge devices. We propose LightEdu-Net, a sensor-noise-adaptive Transformer-based multimodal network that integrates visual, physiological, and environmental signals in a unified lightweight architecture. The model incorporates three key components: a sensor noise adaptive module (SNAM) to suppress degraded sensor inputs, a cross-modal attention fusion module (CMAF) to capture complementary temporal dependencies across modalities, and an edge-aware knowledge distillation module (EAKD) to transfer knowledge from high-capacity teachers to an embedded-friendly student network. We construct a multimodal behavioral dataset from several rural schools and formulate student-state recognition as a multimodal classification task with explicit evaluation of noise robustness and edge deployability. Experiments show that LightEdu-Net achieves 92.4% accuracy with an F1-score of 91.4%, outperforming representative lightweight CNN and Transformer baselines. Under a noise level of 0.3, accuracy drops by only 1.1%, indicating strong robustness to sensor degradation. Deployment experiments further show that the model operates in real time on Jetson Nano with a latency of 42.8 ms (23.4 FPS) and maintains stable high accuracy on Raspberry Pi 4B and Intel NUC platforms. Beyond technical performance, the proposed system provides a low-cost and quantifiable mechanism for capturing fine-grained learning process indicators, offering new data support for educational economics studies on instructional efficiency and resource allocation in underdeveloped regions.

Keywords: LightEdu-Net; state recognition; multimodal sensor fusion; edge-based educational perception; educational economics

1. Introduction

Against the backdrop of rapidly advancing digital education, classroom instruction is moving from experience-driven practices toward data-driven and intelligent decision-making paradigms [1]. Real-time monitoring of student attention, fatigue, and classroom engagement can help teachers track learning dynamics, adjust instructional strategies, and provide an empirical basis for personalized support and resource allocation [2]. However, in rural and under-resourced areas, persistent issues such as insufficient teaching staff, limited classroom supervision, and delayed instructional feedback make it difficult to

obtain such fine-grained behavioral information [3]. Although multimodal sensing and artificial intelligence techniques now enable automated analysis of student states from visual, physiological, and environmental data [4], most existing systems assume reliable sensors, stable networks, and cloud- or GPU-based computation, which rarely hold in these settings. Low-cost devices in real classrooms must cope with severe sensor noise, heterogeneous sampling rates, and strict constraints on computation, memory, and energy. As a result, intelligent tutoring systems, while theoretically valuable for large classes and high teaching workloads [5], remain hard to deploy in the very schools that would benefit most from them. This gap between methodological advances and deployable solutions motivates the design of LightEdu-Net, a multimodal perception framework explicitly tailored to noisy sensors and resource-constrained edge hardware in real rural classrooms.

Although intelligent tutoring systems have achieved notable methodological progress, many existing approaches remain difficult to apply in real classrooms, especially in rural schools [6]. Early work on student behavior monitoring typically relied on single-modal signals (e.g., camera images or wearable physiological data) and traditional machine learning models such as SVM, random forest, and KNN operating on handcrafted features [7,8]. These shallow representations struggle to capture the complexity and variability of classroom environments and are highly sensitive to illumination changes, occlusions, and sensor noise, leading to poor robustness in deployment [9,10]. Moreover, low-dimensional features are inadequate for describing temporal dynamics such as attention fluctuation and cumulative fatigue [11], and most pipelines assume desktop or server-side processing, which prevents real-time inference on the low-cost, low-power devices typically available in rural settings [12].

With the advancement of deep learning, student behaviors have increasingly been modeled using convolutional neural networks, Transformer architectures, and graph neural networks [13]. CNNs and Transformers can extract rich spatial–temporal features from visual, speech, and physiological signals, and multimodal fusion strategies such as early fusion, late fusion, and cross-attention improve recognition accuracy by exploiting complementary information across modalities [14,15]. However, most of these models are computationally heavy, rely on high-performance GPUs or cloud infrastructure, and assume relatively clean and homogeneous multimodal data [16,17]. In contrast, sensors in rural classrooms often produce noisy, low-fidelity, and asynchronously sampled signals, and cloud-based solutions may be unreliable due to limited bandwidth and raise privacy concerns [18]. Recent edge-oriented multimodal frameworks such as TinyM^{2}Net, FarmLightSeek, EduVQA, and lightweight maritime recognition systems demonstrate that compact models can be deployed on resource-constrained hardware in other domains [19,20], but there is still a lack of designs that jointly address sensor noise, heterogeneous educational modalities, and strict edge-computing constraints in real classrooms. This gap motivates our development of LightEdu-Net.

In this study, we focus on how to perform reliable multimodal student-state recognition in rural classrooms where sensing is noisy and computation must run on low-power edge devices. To address this problem, we propose LightEdu-Net, a lightweight multimodal framework designed to be both noise-resilient and deployable. The framework integrates hybrid feature encoding, cross-modal attention, and edge-aware optimization, enabling high-accuracy and low-latency inference on embedded platforms. The main contributions of this work are threefold.

1. We construct a multimodal behavioral dataset from rural schools, capturing visual, physiological, and environmental signals under real classroom conditions. The dataset explicitly contains noise artifacts, illumination changes, and motion disturbances, providing realistic supervision for evaluating robustness in resource-constrained environments.
2. We design LightEdu-Net, a sensor-noise adaptive multimodal network that combines a lightweight convolution–Transformer backbone with a sensor noise adaptive module (SNAM) and a cross-modal attention fusion module (CMAF). This architecture suppresses abnormal sensor inputs, aligns heterogeneous modalities, and extracts discriminative temporal–spatial features for student-state recognition.
3. We validate the framework through targeted experiments that reflect the core challenge: robustness and deployability. Results show that LightEdu-Net maintains high accuracy under increasing noise levels and operates in real time on edge devices such as Jetson Nano, Raspberry Pi 4B, and Intel NUC, demonstrating its suitability for practical deployment in resource-constrained educational settings.
4. From an educational economics perspective, the proposed framework provides low-cost and quantifiable learning process indicators for resource-constrained classrooms, enabling data-driven analysis of teaching efficiency and educational resource utilization in underdeveloped regions.

The remainder of this manuscript is organized as follows. Section 2 reviews related work on intelligent education, multimodal sensor fusion, edge computing, and the educational economics perspective that motivates our focus on rural and resource-constrained classrooms. Section 3 describes the materials and methods, including the construction of the multimodal educational behavior dataset, data preprocessing and augmentation strategies, and the proposed LightEdu-Net framework with its sensor noise adaptive module, cross-modal attention fusion module, and edge-aware knowledge distillation module. Section 4 presents the experimental setup and results, comparing LightEdu-Net with representative baselines, reporting ablation studies and noise-robustness analyses, and evaluating deployment performance on different edge devices, followed by a discussion of practical implications and remaining limitations. Section 5 concludes the paper and outlines directions for future research.

2. Related Work

2.1. Applications of Deep Learning in Intelligent Education

To understand why reliable multimodal perception remains challenging in real classrooms, we first review how deep learning has been applied in intelligent education. The application of deep learning in this field primarily relies on convolutional neural networks, Transformer architectures, and graph neural networks, each modeling student behavior from spatial, temporal, or relational perspectives [21,22]. Convolutional neural networks extract posture variations, facial expression details, and behavioral patterns from video frames through local receptive fields and hierarchical feature mechanisms, offering stable performance in attention monitoring, classroom behavior recognition, and emotion analysis [23]. With the emergence of attention mechanisms, Transformer models leverage global dependency modeling to capture dynamic behavioral changes over time and are particularly effective for states with strong temporal characteristics, such as attention fluctuation and cumulative fatigue [24]. Graph neural networks construct student–student or student–environment relational structures, enabling modeling of pose keypoint relationships and group learning behaviors [25].

Although these deep architectures significantly improve accuracy in educational behavior analysis, their limitations remain notable for deployment in resource-constrained schools. First, their large parameter sizes and high computational complexity generally require GPU acceleration to achieve real-time inference [26]. Second, many of these models are trained on controlled or high-quality datasets and therefore struggle to generalize to rural classrooms characterized by illumination variation, occlusions, missing data, and noise [27]. Third, incorporating multimodal signals further increases architectural complexity, making it difficult for such models to execute smoothly on embedded devices [28]. These constraints hinder the large-scale adoption of deep learning technologies in the very environments that most need support and motivate the lightweight, noise-resilient multimodal modeling focus of this study.

2.2. Multimodal Sensor Fusion Techniques

Reliable student-state recognition in real classrooms often requires combining multiple noisy and heterogeneous signals, making multimodal sensor fusion a central topic in intelligent education research. The primary objective is to integrate visual, physiological, speech, and environmental signals to obtain more comprehensive and stable representations than those provided by any single modality [29]. Early fusion techniques concatenate raw signals or low-level features at the input stage to exploit cross-modal complementarity [30], whereas late fusion aggregates decisions from independently encoded modalities and offers structural simplicity and interpretability [31]. More recently, cross-modal attention mechanisms have been employed to compute mutual information among modalities and adaptively adjust fusion weights, enabling the capture of complex inter-modal relationships in dynamic classroom environments [32]. These methods generally aim to enhance recognition accuracy and robustness [33].

However, when moving from controlled lab settings to rural classrooms, modality heterogeneity, inconsistent sampling frequencies, missing data, and diverse noise sources significantly increase the difficulty of fusion [34]. Simple concatenation or fixed fusion rules are often brittle under such conditions, and even attention-based approaches may waste capacity on corrupted channels if noise is not explicitly modeled. Effective multimodal fusion therefore requires not only expressive feature representations but also mechanisms that remain stable under varying data quality. This motivates the cross-modal attention fusion in LightEdu-Net, which is coupled with sensor-noise adaptation to dynamically adjust fusion strategies when confronted with low-quality, noisy, or heterogeneous classroom data.

2.3. Edge Computing and Lightweight Model Research

With the growing adoption of edge computing, many studies have sought to deploy deep learning models on embedded or low-power devices to reduce cloud dependence, improve real-time performance, and enhance data privacy protection [35]. The core idea of edge computing is to relocate part or all of the inference workload to devices close to the data source, thereby reducing communication overhead and increasing system autonomy [36]. However, the computational and storage requirements of most deep models substantially exceed the capacity of typical edge hardware, prompting the development of model compression and lightweight design techniques [37]. Pruning removes redundant weights to decrease computation, quantization reduces numerical precision to save memory, knowledge distillation transfers information from a large teacher model to a compact student model, and lightweight architectures (e.g., MobileNet, ShuffleNet) optimize structural efficiency by design [38]. These techniques have demonstrated strong performance in image classification, speech recognition, and various edge AI applications such as medical monitoring and traffic management [39].

Nevertheless, multimodal perception tasks are inherently more complex than unimodal tasks, and input heterogeneity often causes compressed models to struggle in maintaining both accuracy and real-time performance on constrained devices. Moreover, most existing lightweight approaches do not explicitly account for the noisy, distribution-unstable, and modality-inconsistent characteristics of educational environments, leading to potential instability when deployed in real classrooms. This gap motivates the edge-aware design in LightEdu-Net, where knowledge distillation and architecture choices are tailored to multimodal, noise-prone student-state recognition on low-power hardware.

2.4. Educational Economics Perspectives on Intelligent Tutoring Systems

Multimodal intelligent tutoring systems demonstrate substantial potential from an educational economics perspective, particularly in improving instructional quality and promoting educational equity in resource-constrained settings [40]. Rural schools commonly face systemic challenges such as insufficient teaching staff, oversized classes, and delayed instructional feedback, which reduce the effectiveness of existing human and material inputs [41]. Lightweight models that can be deployed on edge devices without reliance on high-bandwidth networks or high-performance computing facilities provide timely feedback for teachers and effectively enhance the utilization efficiency of scarce classroom resources [42]. From a cost–effectiveness standpoint, such technologies require relatively modest hardware investments compared to server- or cloud-based solutions, while the associated performance gains can exceed those of many traditional pedagogical interventions, making them particularly attractive for low-resource environments [43]. Consequently, intelligent tutoring systems are not only technological innovations within computer science but also serve as new measurement instruments and intervention tools for educational economics, facilitating more equitable educational opportunities and supporting evidence-based optimization of educational resource allocation across regions [44]. In this context, a noise-resilient, edge-deployable framework such as LightEdu-Net is especially relevant, as it is explicitly designed to operate under the economic and infrastructural constraints that characterize under-resourced schools.

3. Materials and Method

3.1. Data Collection

The data collection process was conducted from March 2024 to January 2025 across authentic classroom environments in three rural primary schools and one county-level middle school in western China, as shown in Table 1 and Figure 1. A total of 12 multimodal acquisition terminals were deployed, integrating high-definition cameras, wearable physiological sensors, and environmental monitoring units. This setting reflects the conditions under which LightEdu-Net is expected to operate: highly heterogeneous signals, frequent illumination changes, and unstable network connectivity. Visual data were captured using 1080p cameras mounted at the front and sides of classrooms at 30 FPS, providing continuous observation of posture, facial expressions, and gaze direction. Automatic exposure control and illumination compensation were enabled to cope with alternating natural and artificial light. Each 40-min lesson yielded approximately 72,000 sequential frames per camera, and infrared auxiliary lighting was used to improve face and keypoint visibility under dim or reflective conditions.

In total, data were collected from 268 students across Grades 3–9, with class sizes ranging from 32 to 48 students. The cohort includes both boys and girls and reflects the typical demographic composition of rural schools in western China in terms of age distribution and socio-economic background. Classroom subjects covered Chinese language, mathematics, English, and science, as well as a smaller number of sessions in moral education and

integrated practical activities, thereby capturing a variety of instructional formats (lecture, discussion, group work, and individual practice).

Physiological signals were recorded through multi-channel wearable devices attached to students' wrists or arms, providing heart rate (HR), electrodermal activity (EDA), temperature (TEMP), and blood oxygen saturation (SpO_2) at 10 Hz. Data were streamed via Bluetooth 5.0 to a local edge node and synchronized with visual data. A caching and resume mechanism was used to address intermittent intranet connectivity, ensuring completeness. To reduce inter-individual variability, physiological signals were normalized and smoothed before model input. Environmental conditions were monitored at 1-second intervals using sensor modules placed at the front, rear, and window areas of the classroom. Measured variables included illumination (lux), noise level (dB), temperature (°C), and humidity (RH). These signals capture rapid fluctuations associated with instructional activities, such as reading aloud, classroom transitions, or external disturbances.

All data collection procedures were approved by the institutional ethics committee of the collaborating university and the local education authorities. Written informed consent was obtained from school administrators, teachers, and the legal guardians of all participating students, and assent was obtained from the students themselves. During acquisition and subsequent processing, personal identifiers were removed or anonymized, and only aggregated indicators are used in analysis and system feedback to protect privacy.

Table 1. Statistics of the multimodal educational behavior dataset.

Data Type	Source	Sampling Rate/FPS	Sample Count
Visual Data	Classroom fixed cameras	30 FPS	2,160,000 frames
Physiological Data	Wearable sensors (HR, EDA, TEMP, SpO_2)	10 Hz	360,000 sets
Environmental Data	LUX/noise/temperature/humidity sensors	1 Hz	540,000 entries

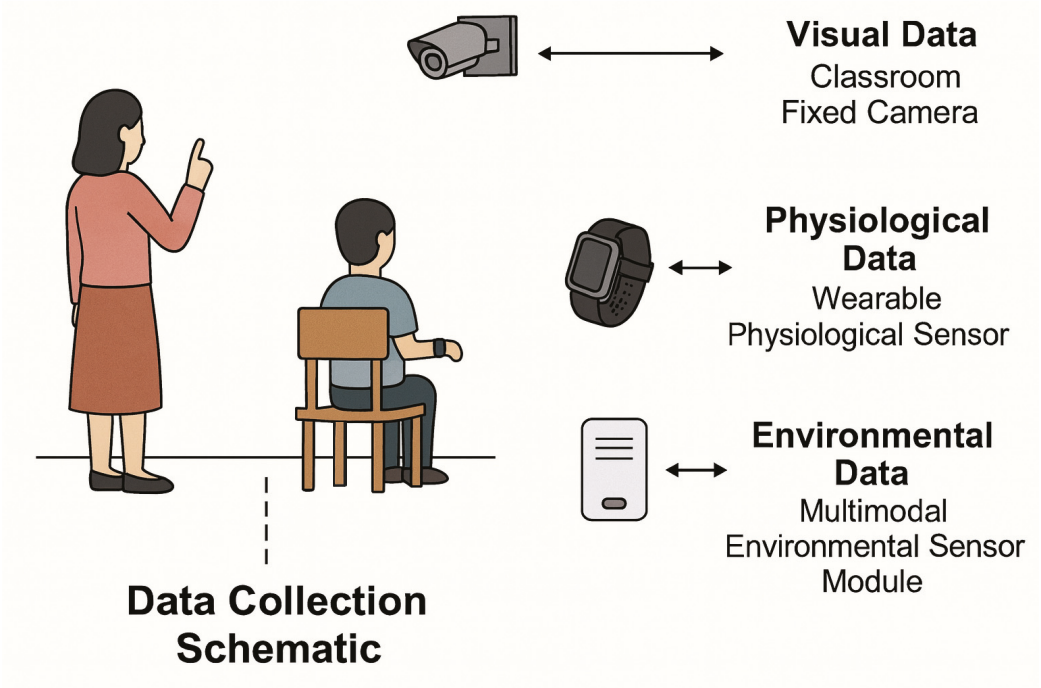


Figure 1. Overview of the multimodal data collection framework used in this study.

3.2. Dataset Processing

Reliable multimodal modeling requires that input signals exhibit comparable scale, temporal consistency, and noise stability. Classroom environments inherently contain illumination fluctuation, partial occlusions, loose-wearing artifacts, and motion-induced perturbations. To ensure that the downstream lightweight architecture receives stable inputs, we apply a hybrid processing strategy that combines basic signal filtering and normalization with targeted augmentation. These operations are not intended as contribution points, but rather provide necessary signal conditioning for the subsequent SNAM and CMAF modules to operate effectively under resource constraints.

3.2.1. Time-Series Data

For physiological signals, filtering represents the first step to suppress power-line interference and high-frequency noise. Using linear time-invariant filtering expressed in convolutional form, the filtered signal is given by

$$y(t) = \int_{-\infty}^{+\infty} x(\tau)h(t-\tau)\,\mathrm{d}\tau, \tag{1}$$

where $x(t)$ is the raw signal and $h(t)$ the impulse response of the filter. In practice, band-pass filters retain cognitive-related frequencies (e.g., 0.1–40 Hz) while removing out-of-band disturbances commonly observed in wearable classroom measurements.

Since inconsistent magnitudes across modalities may disturb optimization, all physiological and visual features are standardized. The standardized signal is defined as

$$\hat{x} = \frac{x-\mu}{\sigma}, \tag{2}$$

where μ and σ denote mean and standard deviation. To accommodate non-stationary statistics in long classroom recordings, a sliding-window standardization strategy is adopted:

$$\hat{x}_t = \frac{x_t - \mu_t^{(W)}}{\sigma_t^{(W)}}, \tag{3}$$

where $\mu_t^{(W)}$ and $\sigma_t^{(W)}$ are the window-based mean and variance computed within W samples.

Sudden spikes caused by electrode displacement or motion artifacts are further smoothed by windowed averaging,

$$\tilde{x}_t = \frac{1}{|W|}\sum_{i=t-k}^{t+k} x_i, \tag{4}$$

where $|W| = 2k+1$. This preserves global trends while attenuating high-frequency jitter, producing temporally coherent features that SNAM later refines via learnable denoising.

3.2.2. Image Data

Visual signals are augmented to simulate realistic disturbances. Brightness perturbation is modeled by a linear intensity transformation:

$$I' = \gamma I + \delta, \tag{5}$$

where γ scales global brightness and δ models ambient light shift. This augmentation improves robustness to lighting variance often observed in classrooms with windows or ceiling lights.

Occlusion is simulated using a Cutout strategy. A random rectangular region Ω is selected and replaced:

$$I'(x,y) = \begin{cases} 0, & (x,y) \in \Omega, \\ I(x,y), & \text{otherwise.} \end{cases} \tag{6}$$

This encourages the backbone to rely on multiple cues rather than single facial or postural regions, improving resilience when students move, lean, or are partially blocked.

3.3. Data Enhancement

To reduce over-sensitivity to sampling jitter, a temporal perturbation strategy is introduced. For each time step, a small offset ϵ_t is added:

$$x'_t = x_{t+\epsilon_t}, \tag{7}$$

where ϵ_t follows a zero-mean, small-variance distribution. Such temporal jittering prevents overfitting to rigid time alignment and simulates realistic sensor drift.

Multimodal signals often differ in sampling rate. To allow cross-modal attention to operate on a unified temporal axis, all modalities are resampled via interpolation. For an original signal $x(t)$ and resampled timeline t', we compute

$$x'(t') = \text{Interp}(x(t), t'), \tag{8}$$

where $\text{Interp}(\cdot)$ denotes linear interpolation. This temporal synchronization ensures that CMAF processes inputs within consistent windows, and that modality interactions reflect true behavioral dynamics rather than sampling discrepancies.

Taken together, these preprocessing steps provide stable and comparable multimodal inputs. They do not constitute novel algorithmic contributions, but they are essential to enable SNAM to learn frequency-domain denoising and CMAF to exploit cross-modal mutual information under noisy classroom conditions and edge-computing resource limits.

3.4. Proposed Method

3.4.1. Overall

We propose LightEdu-Net, a noise-resilient multimodal student-state recognition framework designed for edge deployment in real classrooms. The model takes temporally aligned visual, physiological, and environmental features as input and passes them through three lightweight modality-specific encoders: a convolution–Transformer hybrid backbone for visual signals, a 1D convolutional–recurrent backbone for physiological signals, and multi-channel convolutions for environmental measurements. The encoded features are then refined by the sensor noise adaptive module (SNAM), which performs learnable band-pass denoising and channel reweighting to improve the signal-to-noise ratio before fusion. The denoised modality features are subsequently fed into the cross-modal attention fusion (CMAF) layer, where intra-modal attention emphasizes salient temporal segments and inter-modal attention captures complementary dependencies across modalities, yielding a compact joint representation. On top of this fused representation, a lightweight prediction head produces student-state labels suitable for classroom monitoring on embedded devices. During training, an edge-aware knowledge distillation (EAKD) scheme is applied, in which a high-capacity teacher network guides the LightEdu-Net student to approximate its decision boundaries while keeping the student architecture small and quantization-friendly. At inference time, only the student network is used: data flow sequentially through the encoders, SNAM, CMAF, and the prediction head, enabling robust multimodal perception with low latency on low-power edge hardware.

3.4.2. Sensor Noise Adaptive Module

The sensor noise adaptive module (SNAM) is inserted between the modality-specific encoders and the multimodal fusion layer, following the noise-rectification pipeline illustrated in Figure 2. For each modality, the encoder output is first projected to a compact temporal feature sequence. Concretely, the visual backbone produces a tensor $\mathbf{V} \in \mathbb{R}^{T\times 14\times 14\times 64}$, which is passed through a 1×1 pointwise convolution (32 output channels) and global spatial average pooling to obtain $\tilde{\mathbf{V}} \in \mathbb{R}^{T\times 32}$. The physiological branch yields $\tilde{\mathbf{P}} \in \mathbb{R}^{T\times 64}$ through one-dimensional convolutions and GRU layers, and the environmental branch produces $\tilde{\mathbf{E}} \in \mathbb{R}^{T\times 32}$ via multi-channel convolutions. SNAM processes each modality feature sequence $X \in \{\tilde{\mathbf{V}}, \tilde{\mathbf{P}}, \tilde{\mathbf{E}}\}$ independently with shared parameters, so that the same denoising mechanism is applied to visual, physiological, and environmental signals.

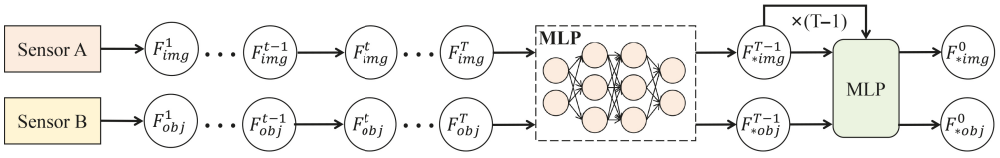


Figure 2. This figure illustrates the overall workflow of the sensor noise adaptive module (SNAM).

For a given modality, we first apply a lightweight temporal encoder to capture local dynamics and smooth high-frequency fluctuations. A depthwise–separable temporal convolution stack DS−TCN is used,

$$\mathbf{Z} = \mathrm{DS{-}TCN}(X), \tag{9}$$

where two depthwise–separable Conv1d layers (kernel sizes $k = 5$ and $k = 3$, dilations $d = 1$ and $d = 2$) with LayerNorm and SiLU activation preserve the shape $\mathbf{Z} \in \mathbb{R}^{T\times C}$ (with $C \in \{32, 64\}$). To further suppress noise in the spectral domain, we perform a learnable band-pass filtering. Let $\mathcal{F}_t$ denote the FFT along the temporal axis and $\hat{\mathbf{Z}} = \mathcal{F}_t(\mathbf{Z})$ the corresponding spectrum. We define a parameterized mask

$$M(\omega) = \sigma\big(s(\omega - \omega_{\mathrm{low}})\big)\,\sigma\big(s(\omega_{\mathrm{high}} - \omega)\big), \tag{10}$$

where ω_{low} and ω_{high} are learnable cutoff frequencies, $s > 0$ is a learnable slope, and $\sigma(\cdot)$ denotes the sigmoid function. The filtered features in the time domain are then obtained by

$$\mathbf{U} = \mathcal{F}_t^{-1}\big(M \odot \hat{\mathbf{Z}}\big), \tag{11}$$

which attenuates frequency components outside the band of interest while preserving the temporal resolution.

Residual noise is often concentrated in a subset of channels, especially under sensor jitter or loose wearing. To adaptively down-weight such channels, we introduce a channel reweighting unit. Global temporal average pooling GAP_t followed by two bottleneck fully connected layers produces a set of channel importance scores $\mathbf{w} \in \mathbb{R}^C$:

$$\mathbf{w} = \sigma\Big(\mathbf{W}_2\,\phi\big(\mathbf{W}_1\,\mathrm{GAP}_t(\mathbf{U})\big)\Big), \quad \mathbf{W}_1 \in \mathbb{R}^{\frac{C}{r}\times C},\ \mathbf{W}_2 \in \mathbb{R}^{C\times\frac{C}{r}}, \tag{12}$$

where $\phi(\cdot)$ is the SiLU activation and $r = 8$ is the reduction ratio. The final SNAM output is

$$\mathbf{Y} = \mathbf{U} \odot \mathbf{w}, \tag{13}$$

which preserves the temporal dimension T and channel dimension C of the input. Channels dominated by noise receive smaller weights, whereas informative channels are retained.

To stabilize training and encourage meaningful masks, we introduce two auxiliary objectives. A simple spectral regularization penalizes excessive suppression of useful bands,

$$\mathcal{L}_{\text{band}} = \left\| (1 - M) \odot \hat{\mathbf{Z}} \right\|_2^2, \tag{14}$$

where M is the learnable frequency mask and $\hat{\mathbf{Z}}$ is the temporal spectrum of the intermediate features. In addition, when teacher features or clean augmented views are available, we encourage the denoised features to stay close to a reference representation. Let $\mathbf{Y} \in \mathbb{R}^{T \times C}$ denote the SNAM output for the current modality, where $\mathbf{Y}_{t,c}$ is the value at time step t and channel c, and let $T \in \mathbb{R}^{T \times C}$ denote the corresponding teacher features (or clean augmented views) extracted at the same layer and temporal resolution, with $T_{t,c}$ aligned to the same position. The alignment loss is then defined as

$$\mathcal{L}_{\text{align}} = \frac{1}{TC} \sum_{t,c} \left| \mathbf{Y}_{t,c} - T_{t,c} \right|. \tag{15}$$

In the overall LightEdu-Net framework, the modality-wise outputs $\mathbf{Y}^{(v)} \in \mathbb{R}^{T \times 32}$, $\mathbf{Y}^{(p)} \in \mathbb{R}^{T \times 64}$, and $\mathbf{Y}^{(e)} \in \mathbb{R}^{T \times 32}$ are concatenated and projected to a unified embedding dimension before being fed into the cross-modal attention fusion module. In this rectification-before-fusion design, SNAM improves the signal-to-noise characteristics of each modality so that CMAF can allocate its attention capacity to physically meaningful temporal patterns rather than corrupted sensor artifacts, which is particularly important in noisy, resource-constrained classroom environments.

3.4.3. Cross-Modal Attention Fusion

The cross-modal attention fusion (CMAF) module is designed to model interactive representations across modalities rather than within a single modality. Conventional self-attention computes dependencies using query, key, and value matrices (Q, K, V) derived from the same input, and thus primarily captures intra-modal structure. In contrast, CMAF explicitly exchanges information among visual, physiological, and environmental branches, allowing each modality to attend to others in a shared embedding space.

As illustrated in Figure 3, CMAF operates on the SNAM-processed features and consists of stacked Transformer-style blocks. Each block contains intra-modal self-attention, inter-modal cross-attention, residual connections with LayerNorm, and a two-layer multilayer perceptron (MLP). The visual, physiological, and environmental branches receive inputs of shape $T \times C_m$, where T denotes the sequence length and $C_m \in \{32, 64, 32\}$ is the channel dimension for the three modalities, respectively. All modality-specific features are first projected to a unified embedding dimension $d = 96$ so that they can be processed jointly. In our implementation, CMAF uses three such blocks with $h = 8$ attention heads per block and head dimension $d_h = d/h = 12$.

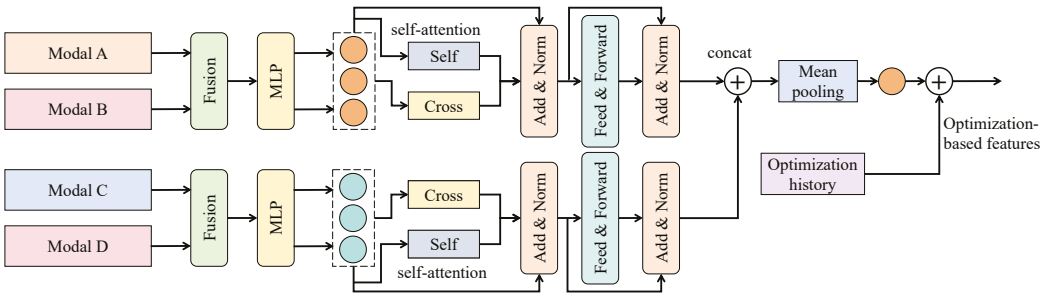


Figure 3. Architecture of the cross-modal attention fusion (CMAF) module.

Formally, let $\mathbf{X}_v, \mathbf{X}_p, \mathbf{X}_e \in \mathbb{R}^{T\times d}$ denote the visual, physiological, and environmental features after linear projection. For notational simplicity, we describe the update of the visual branch using cross-attention to the physiological and environmental branches; the other branches are handled analogously. We define linear projections

$$Q_v = \mathbf{X}_v\mathbf{W}_q, \quad K_p = \mathbf{X}_p\mathbf{W}_k, \quad V_p = \mathbf{X}_p\mathbf{W}_v, \quad K_e = \mathbf{X}_e\mathbf{W}_k, \quad V_e = \mathbf{X}_e\mathbf{W}_v, \tag{16}$$

where $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v \in \mathbb{R}^{d\times d_h}$ are learnable projection matrices shared across modalities within a block. Cross-modal attention from modality p to modality v is computed as

$$\mathbf{A}_{v\leftarrow p} = \mathrm{Softmax}\left(\frac{Q_v K_p^\top}{\sqrt{d_h}}\right)V_p, \tag{17}$$

and similarly for attention from modality e to modality v,

$$\mathbf{A}_{v\leftarrow e} = \mathrm{Softmax}\left(\frac{Q_v K_e^\top}{\sqrt{d_h}}\right)V_e. \tag{18}$$

The $\mathrm{Softmax}(\cdot)$ operator is applied row-wise over the temporal dimension to yield attention weights for each time step. The updated visual representation is given by

$$\mathbf{Y}_v = \mathbf{X}_v + \alpha_p\mathbf{A}_{v\leftarrow p} + \alpha_e\mathbf{A}_{v\leftarrow e}, \tag{19}$$

where α_p and α_e are learnable scalar coefficients satisfying $\alpha_p + \alpha_e = 1$ and controlling the relative contribution of physiological and environmental information. In this way, the visual branch can adaptively integrate cues such as physiological rhythm changes and environmental disturbances at each time step. The same procedure is used to update $\mathbf{Y}_p$ and $\mathbf{Y}_e$ by exchanging roles of the modalities. After normalization and concatenation along the channel axis, we obtain a fused representation $\mathbf{F} \in \mathbb{R}^{T\times 3d}$, which is then projected back to $T \times d$ through an MLP and forwarded to the final prediction head.

From an information-theoretic perspective, CMAF is designed to increase the mutual information among modalities. Let $I(X;Y)$ denote the mutual information between two random variables. While single-modal self-attention mainly enhances $I(X_v;X_v)$, $I(X_p;X_p)$, and $I(X_e;X_e)$, CMAF aims to strengthen cross-modal terms such as $I(X_v;X_p)$, $I(X_v;X_e)$, and $I(X_p;X_e)$. A common variational lower bound for $I(X_i;X_j)$ can be written as

$$I(X_i;X_j) \geq \mathbb{E}_{(x_i,x_j)}\big[\log D_\theta(x_i,x_j)\big] - \log \mathbb{E}_{x_i}\mathbb{E}_{x_j}\big[\exp(D_\theta(x_i,x_j))\big], \tag{20}$$

where $D_\theta(\cdot,\cdot)$ is a learnable critic function. In our setting, the attention score matrices implicitly play a similar role to D_θ by assigning higher weights to modality pairs that are more informative for the task, thereby encouraging the network to align semantically consistent patterns across modalities.

When combined with the preceding SNAM module, CMAF operates in a rectification-before-alignment manner. SNAM first enhances the signal-to-noise ratio of each modality and suppresses corrupted channels, and CMAF then performs cross-modal alignment and fusion on these rectified features. Together with the edge-aware knowledge distillation mechanism, which regularizes the fused representations via a high-capacity teacher network, this hierarchical design yields a robust semantic-fusion core that can handle noisy, asynchronous, and heterogeneous multimodal signals in resource-constrained educational environments.

3.4.4. Edge-Aware Knowledge Distillation

The edge-aware knowledge distillation (EAKD) module aims to balance lightweight feature extraction and high-accuracy prediction on edge devices by introducing a dual-teacher distillation strategy. In this way, the lightweight student network deployed on devices such as Jetson Nano and Raspberry Pi can inherit the decision capability of more expressive teacher models trained on high-performance servers. As illustrated in Figure 4, EAKD involves two teachers and one student: a CNN teacher (T_{CNN}), a ViT teacher, and a lightweight CNN student (denoted as S_{CNN}). The teacher models possess stronger local and global modeling capacity, whereas the student model contains only essential convolutional and attention layers and is designed to meet edge-computing constraints.

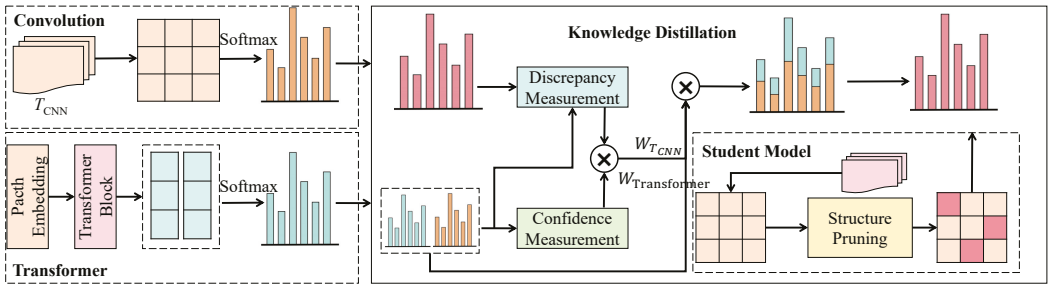


Figure 4. Illustration of the edge-aware knowledge distillation (EAKD) module.

To adaptively combine the two teachers, we adopt a discrepancy-aware teacher weighting (DATW) scheme. For each teacher $k \in \{\text{CNN}, \text{ViT}\}$, we define a confidence score

$$C_k(x) = \max_c \mathbf{p}_{T_k,c}(x), \tag{21}$$

that reflects how certain the teacher is about its prediction, and a discrepancy measure

$$D_k(x) = \left\| \mathbf{z}_{T_k}(x) - \mathbf{z}_S(x) \right\|_2, \tag{22}$$

that reflects how far the student currently deviates from that teacher in logit space. The unnormalized teacher weight is then given by

$$\tilde{\omega}_{T_k}(x) = \frac{C_k(x)}{D_k(x) + \epsilon}, \tag{23}$$

where $\epsilon > 0$ is a small constant to avoid division by zero. The normalized weights are obtained as

$$\omega_{T_k}(x) = \frac{\tilde{\omega}_{T_k}(x)}{\tilde{\omega}_{T_{\text{CNN}}}(x) + \tilde{\omega}_{T_{\text{ViT}}}(x)}, \tag{24}$$

so that $\omega_{T_{\text{CNN}}}(x) + \omega_{T_{\text{ViT}}}(x) = 1$ for all x. The combined teacher target, illustrated as $Z_{target}(x)$ in Figure 4, is a convex combination of the two teacher logits:

$$\mathbf{z}_{\text{target}}(x) = \omega_{T_{\text{CNN}}}(x)\, \mathbf{z}_{T_{\text{CNN}}}(x) + \omega_{T_{\text{ViT}}}(x)\, \mathbf{z}_{T_{\text{ViT}}}(x). \tag{25}$$

The EAKD loss integrates logit matching and distribution-level distillation. The logit-level term encourages the student to approximate the combined teacher target $Z_{target}(x)$,

$$\mathcal{L}_{\text{logit}} = \left\| \mathbf{z}_S(x) - \mathbf{z}_{\text{target}}(x) \right\|_2^2, \tag{26}$$

and the distribution-level term uses Kullback–Leibler divergence to pull the student probabilities toward the teacher probabilities,

$$\mathcal{L}_{\mathrm{KL}} = \sum_{k \in \{\mathrm{CNN,ViT}\}} \mathrm{KL}\big(\mathbf{p}_{T_k}(x) \,\big\|\, \mathbf{p}_S(x)\big). \tag{27}$$

The overall EAKD objective is then written as

$$\mathcal{L}_{\mathrm{EAKD}} = \eta\, \mathcal{L}_{\mathrm{logit}} + (1-\eta)\, \mathcal{L}_{\mathrm{KL}}, \tag{28}$$

where $\eta \in [0,1]$ is a balancing coefficient tuned on the validation set. When the student already matches a given teacher well (small $D_k(x)$), the corresponding weight $\omega_{T_k}(x)$ increases, and that teacher dominates the distillation signal; when the discrepancy is large or the teacher is uncertain (low $C_k(x)$), its influence is automatically reduced.

By combining structurally different teachers (CNN and ViT), explicitly modeling teacher confidence and student–teacher discrepancy, and constraining the student to a lightweight architecture, EAKD enables LightEdu-Net to approach the decision quality of high-capacity models while maintaining low computational cost. In the intelligent tutoring task considered here, this module increases recognition accuracy on low-power platforms, reduces inference latency, and alleviates dependence on high-bandwidth transmission, thereby achieving a favorable balance between computation and intelligence for edge-deployed educational models.

4. Results and Discussion

4.1. Experimental Settings

4.1.1. Platforms and Parameters

The experimental hardware platforms employed in this study encompass three typical categories of edge and near-edge computing devices to comprehensively validate the model's adaptability and performance under varying computational capacities. NVIDIA Jetson Nano, equipped with a 128-core Maxwell GPU, serves as a representative low-power embedded platform commonly available in rural schools. Raspberry Pi 4B contains a quad-core ARM Cortex-A72 CPU and performs inference purely on processor resources, aligning with extremely low-compute deployment scenarios. Intel NUC, as a lightweight desktop-class system, integrates a multi-core CPU with higher clock frequency, enabling evaluation of real-time performance and energy consumption under moderate computational conditions. Deployment across this hardware spectrum provides a realistic demonstration of the computational heterogeneity and operational demands likely encountered in real-world educational environments.

Regarding software configuration, the experiments are conducted using PyTorch 2.2 as the primary deep learning framework for model construction and training. GPU-accelerated inference is enabled through CUDA 12.1 to ensure efficient computation for convolutional and Transformer-based operations. Python 3.10 is adopted as the primary programming language to support data preprocessing, model training, visualization, and general experimental workflow management. All components are executed within an Ubuntu environment to ensure dependency compatibility and stable model deployment, thereby facilitating consistent migration to edge devices. For hyperparameter configuration, the constructed dataset is partitioned into training, validation, and test sets with a ratio of 7:2:1, ensuring adequate multimodal feature learning during training while preserving sufficient samples for performance assessment. A mini-batch strategy with batch size = 32 is adopted to balance memory consumption and stable gradient updates. The learning rate is set to 1×10^{-4} and optimized with the Adam optimizer to achieve adaptive gradient

updates while maintaining stable convergence. A total of 100 training epochs is used to allow sufficient convergence time. To further evaluate the generalization capability under multimodal input conditions, a five-fold cross-validation strategy is incorporated. Specifically, the dataset is randomly divided into five subsets, with four subsets used for training and one for validation in each iteration. After five iterations, the averaged performance metrics are obtained, reducing bias introduced by data partitioning and yielding a more reliable estimation of model performance.

4.1.2. Baseline Models and Evaluation Metrics

A set of representative lightweight and multimodal baseline models is selected for comparison, including MobileNetV3 + SVM [45,46], ResNet [47], EfficientNet-lite [48], Tiny-Transformer [49], and CLIP-Fusion [50]. MobileNetV3 + SVM combines an efficient feature extraction backbone with a traditional classifier, offering advantages in structural simplicity and training stability. ResNet, as a classical residual architecture, provides strong robustness and consistent performance. EfficientNet-lite adopts a compound scaling strategy across depth, width, and resolution, achieving high parameter efficiency with relatively strong accuracy. Tiny-Transformer employs a lightweight self-attention structure by reducing attention heads and channel dimensions, enabling long-range dependency modeling under constrained computational budgets. CLIP-Fusion exploits pretrained vision–language alignment models as multimodal encoders, enabling feature concatenation or matching across modalities and providing strong cross-modal transferability. For all baselines, network depth, width, learning rate, and batch size are selected according to their original papers or official lightweight configurations and then fine-tuned on a held-out validation set to ensure a fair comparison under similar computational budgets.

The evaluation in this study covers classification metrics (accuracy, precision, recall, F1-score), efficiency metrics (model size, FLOPs, inference latency, power consumption), and robustness metrics (performance degradation under noise, expressed as $\Delta Accuracy$), providing a comprehensive assessment of model suitability in resource-constrained educational environments. The mathematical definitions of the evaluation metrics are as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{29}$$

$$Precision = \frac{TP}{TP + FP}, \tag{30}$$

$$Recall = \frac{TP}{TP + FN}, \tag{31}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}, \tag{32}$$

$$\Delta Accuracy = Accuracy_{clean} - Accuracy_{noise}. \tag{33}$$

The symbols in these equations are defined as follows: TP denotes the number of true positives, representing instances correctly identified as positive; TN denotes true negatives, referring to instances correctly identified as negative; FP denotes false positives, referring to negative instances incorrectly classified as positive; FN denotes false negatives, representing positive instances incorrectly classified as negative. $Accuracy_{clean}$ denotes the accuracy obtained under noise-free conditions, whereas $Accuracy_{noise}$ represents the accuracy under specific noise perturbations, thereby reflecting the model's sensitivity to noise.

4.2. Comparison of LightEdu-Net and Baseline Models

The objective of this experiment is to verify whether the proposed LightEdu-Net can maintain high recognition accuracy while remaining lightweight and computationally efficient in multimodal student-state recognition tasks, thereby satisfying real-time inference demands on resource-constrained edge devices. By comparing LightEdu-Net with five representative lightweight and multimodal baseline models, a comprehensive evaluation of the trade-offs among classification capacity, representational ability, and computational efficiency under different architectural designs can be achieved.

As shown in Table 2 and Figure 5, conventional convolution-based models such as MobileNetV3 + SVM and ResNet18 exhibit considerably lower accuracy and recall. Their feature extraction capacity is restricted by shallow convolutional kernels with limited receptive fields, which are insufficient for capturing long-range temporal dependencies and subtle cross-modal cues, such as gradual attention changes or cumulative physiological fatigue. EfficientNet-lite and Tiny-Transformer provide stronger representational capacity but still fall short in explicitly modeling and exploiting complementary information across visual, physiological, and environmental inputs. CLIP-Fusion, as a large pretrained multimodal model, achieves relatively high accuracy; however, its large parameter scale and high FLOPs make deployment on edge devices impractical in rural classroom settings. In contrast, LightEdu-Net achieves the best performance across all metrics, with particularly notable gains in F1-score and recall, because SNAM improves the signal-to-noise ratio at the feature level, CMAF explicitly aligns and fuses modalities through cross-attention, and EAKD transfers knowledge from high-capacity teachers into a compact student. This combination yields a model that is both more discriminative for complex multimodal patterns and better suited to low-power edge deployment.

Table 2. Comparison of LightEdu-Net and baseline models in classification performance and efficiency.

Model	Accuracy	Precision	Recall	F1-Score	Model Size (MB)	FLOPs (G)
MobileNetV3 + SVM	81.2 ± 0.6%	78.9 ± 0.5%	77.5 ± 0.7%	78.2 ± 0.6%	9.8	0.65
ResNet18	84.5 ± 0.4%	83.2 ± 0.6%	82.7 ± 0.5%	82.9 ± 0.4%	44.7	1.82
EfficientNet-lite	86.1 ± 0.5%	85.4 ± 0.4%	84.2 ± 0.6%	84.7 ± 0.5%	13.5	0.92
Tiny-Transformer	87.3 ± 0.4%	86.1 ± 0.5%	85.9 ± 0.3%	86.0 ± 0.4%	15.8	1.35
CLIP-Fusion	88.5 ± 0.3%	87.4 ± 0.4%	87.1 ± 0.3%	87.2 ± 0.3%	328.3	4.20
LightEdu-Net (Proposed)	92.4 ± 0.2%	91.7 ± 0.3%	91.2 ± 0.2%	91.4 ± 0.3%	34.5	1.05

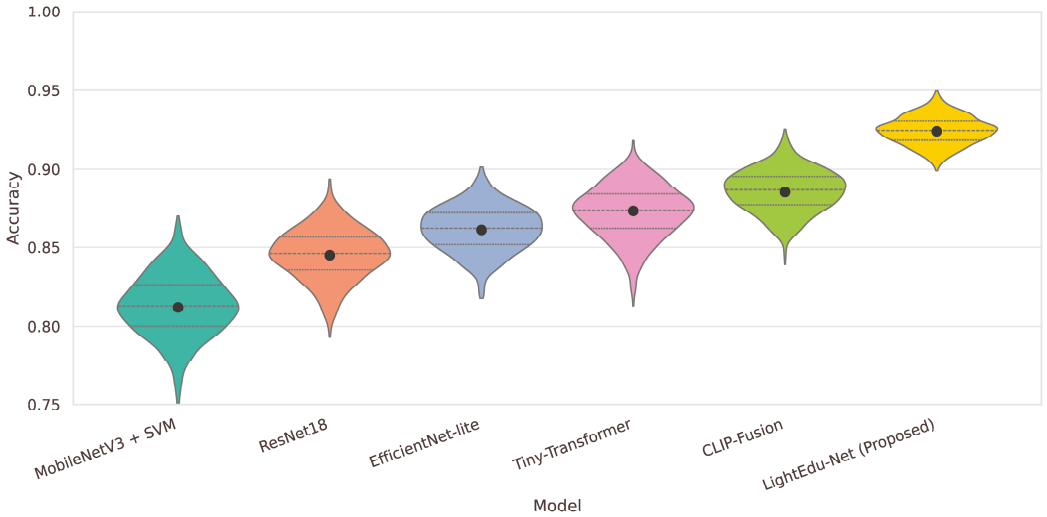


Figure 5. The violin plot of accuracy comparision results across different models.

From a theoretical perspective, performance differences arise from variations in intrinsic structural design and mathematical representational capacity. MobileNetV3 + SVM and ResNet18 rely on local convolutional kernels and lack mechanisms for long-range or multimodal dependency modeling, resulting in representational bottlenecks when handling high-dimensional heterogeneous features from visual, physiological, and environmental signals. EfficientNet-lite increases network capacity through compound scaling, yet it primarily captures patterns within the visual domain, limiting cross-modal fusion ability. Tiny-Transformer can model long-range dependencies but constrains attention computation within the visual modality, preventing deeper multimodal interaction. CLIP-Fusion optimizes a cross-modal embedding space through large-scale pretraining but suffers from high-dimensional embeddings and dense attention structures, resulting in prohibitive computational cost for edge inference. LightEdu-Net achieves superior performance by incorporating a sensor noise adaptive module (SNAM) to enhance feature stability, using a cross-modal attention fusion (CMAF) module to maximize mutual information across modalities, and adopting an edge-aware knowledge distillation (EAKD) mechanism to compress the teacher model's decision boundaries into a lightweight student network. This combination allows the lightweight model to approximate the expressiveness of a high-capacity model while maintaining efficiency, thereby achieving optimal performance under resource constraints.

4.3. Ablation Study

This experiment evaluates whether the overall performance of LightEdu-Net relies on its core modules and quantifies the contributions of each component to multimodal recognition capability and noise robustness.

As shown in Table 3, the three unimodal variants (Visual-only, Physiological-only, and Environmental-only) all perform noticeably worse than the full multimodal LightEdu-Net, indicating that each single modality provides only partial information and that robust student-state recognition benefits from integrating complementary cues. Visual-only typically captures posture and facial dynamics but lacks direct information about internal arousal and external disturbances; Physiological-only is sensitive to fatigue and stress but cannot distinguish specific classroom behaviors; Environmental-only reflects illumination and noise changes but is insufficient for fine-grained engagement recognition. The performance of these variants highlights the necessity of multimodal fusion. retaining only the backbone network leads to a significant decline across all performance metrics, demonstrating that shallow convolutional or limited long-range modeling is insufficient for capturing deep multimodal dependencies. Removing SNAM results in the most severe performance degradation, especially under noise level 0.3, indicating that illumination variations, sensor jitter, and physiological noise—frequently present in real-world rural classrooms—substantially affect feature stability. Removing CMAF moderately reduces performance, confirming that cross-modal attention plays a critical role in exploiting complementary information among modalities. Removing EAKD leads to a mild performance drop, showing that the distillation mechanism effectively enhances the expressive capacity of the lightweight student model without increasing computational cost. From a theoretical standpoint, SNAM stabilizes temporal and spectral characteristics by reducing distributional shifts from noise perturbations. CMAF maximizes mutual information across modalities, equivalent to learning conditional dependency structures in a high-dimensional feature space. Removing CMAF degrades joint representations into weakly coupled features. EAKD provides a richer teacher-defined feature space, allowing the lightweight model to maintain high-level decision boundaries; removing it reduces classification granularity and generalization capability.

Table 3. Ablation study of LightEdu-Net showing the contribution of each module and input configuration.

Model Variant	Accuracy	Precision	Recall	F1-Score	ΔAccuracy (Noise = 0.3)
Visual-only	87.3 ± 0.5%	86.5 ± 0.4%	86.1 ± 0.6%	86.3 ± 0.5%	−8.2%
Physiological-only	84.6 ± 0.6%	83.8 ± 0.5%	83.1 ± 0.7%	83.4 ± 0.6%	−10.1%
Environmental-only	82.9 ± 0.7%	82.1 ± 0.6%	81.4 ± 0.8%	81.7 ± 0.7%	−11.6%
Without SNAM	88.1 ± 0.6%	87.2 ± 0.5%	86.7 ± 0.7%	86.9 ± 0.6%	−7.8%
Without CMAF	89.4 ± 0.5%	88.1 ± 0.4%	87.9 ± 0.6%	88.0 ± 0.5%	−5.2%
Without EAKD	90.2 ± 0.4%	89.5 ± 0.5%	89.1 ± 0.3%	89.2 ± 0.4%	−2.1%
Backbone Only	85.7 ± 0.7%	84.9 ± 0.6%	84.1 ± 0.8%	84.5 ± 0.7%	−10.4%
Full LightEdu-Net	92.4 ± 0.3%	91.7 ± 0.3%	91.2 ± 0.2%	91.4 ± 0.3%	−1.1%

4.4. Deployment Performance

This experiment evaluates the deployability and stability of LightEdu-Net under different computational resources by testing latency, throughput, power consumption, and accuracy on three representative edge and near-edge platforms. The goal is to determine whether the model satisfies real-time, energy-efficient, and reliable recognition requirements in practical educational environments.

As shown in Table 4, Jetson Nano, despite limited computational capability, achieves relatively high throughput and low latency due to the lightweight convolutional architecture and compact fused features, while maintaining server-level accuracy. Raspberry Pi 4B, relying solely on CPU inference, exhibits higher latency but retains stable accuracy under low power consumption, confirming that the model design minimizes convolutional depth and attention complexity to accommodate CPU-based sequential operations. Intel NUC, with higher clock frequencies and multi-core processing, provides the lowest latency and highest throughput, indicating strong scalability on more capable platforms.

Table 4. Deployment performance of LightEdu-Net across different edge and near-edge platforms.

Hardware Platform	Mean Latency (ms)	Throughput (FPS)	Power Consumption (W)	Accuracy
Jetson Nano	42.8	23.4	7.1	92.4%
Raspberry Pi 4B	108.5	9.1	5.3	92.1%
Intel NUC	18.9	52.7	14.5	92.4%

4.5. Discussion

4.5.1. Effectiveness of LightEdu-Net on Core Challenges

The experimental results validate that LightEdu-Net effectively addresses the key challenges of noisy sensing, multimodal fusion, and edge deployability in rural classroom environments. As shown in Table 2 and Figure 5, LightEdu-Net achieves the highest accuracy, precision, recall, and F1-score among all compared methods while maintaining a moderate model size and FLOPs, indicating a favorable balance between representational capacity and computational efficiency. The ablation results in Table 3 further clarify the contribution of each core component: the "Backbone Only" variant exhibits pronounced performance degradation, removing SNAM causes the largest drop and severe sensitivity to noise, removing CMAF weakens the exploitation of complementary multimodal information, and removing EAKD reduces performance despite unchanged inference cost. Together with the deployment results in Table 4, which show that LightEdu-Net meets latency, throughput, and power constraints on Jetson Nano, Raspberry Pi 4B, and Intel NUC, these findings confirm that SNAM, CMAF, and EAKD work synergistically to provide robust, accurate, and deployable multimodal perception under realistic edge-computing conditions.

From an educational perspective, field feedback from teachers suggests that the system's continuous indicators of attention, fatigue, and participation provide actionable

information that is difficult to obtain in large classes, enabling timely adjustments in pacing, interaction strategies, and classroom management. This indicates that LightEdu-Net can help increase the effective use of limited teaching resources and improve instructional efficiency in rural and under-resourced schools. At the same time, the sensitivity of multimodal behavioral and physiological data highlights the need for privacy-preserving extensions, such as federated learning, in which model updates rather than raw data are shared across schools. Future work will also involve more fine-grained power-consumption measurements under varying workloads (e.g., different class sizes and sampling rates) to better assess long-term sustainability. Overall, the current experiments show that the core challenges identified in this study—sensor noise, multimodal fusion, and edge deployment—are substantially mitigated, and that LightEdu-Net provides a practical and scalable foundation for multimodal perception in real classroom environments.

4.5.2. Classroom Usage, Annotation Protocol, and Educational Impact

Beyond quantitative performance, it is important to clarify how LightEdu-Net operates at the behavioral level and how its outputs are obtained. In our implementation, the model predicts discrete engagement-related states, including multi-level attention labels (high/medium/low), fatigue levels (alert/mild/high), and participation indicators (active/passive), derived from the fused multimodal representation produced by CMAF and generated at regular time intervals for each student and the class as a whole. Ground truth labels are constructed through an offline annotation protocol: synchronized video, physiological signals, and classroom logs are reviewed by trained annotators with educational backgrounds, who assign attention, fatigue, and participation levels based on observable cues such as gaze orientation, posture, facial expressions, interaction frequency, and characteristic changes in physiological signals. Each segment is labeled independently by at least two annotators, and disagreements are resolved by consensus, ensuring that the supervisory signals correspond to pedagogically meaningful states rather than purely low-level patterns.

In real deployments, the outputs of LightEdu-Net are delivered to teachers via a lightweight classroom dashboard designed to impose minimal cognitive load. At the individual level, the interface presents smoothed timelines of attention and fatigue using simple color-coded bars without exposing raw physiological values, while at the class level it provides aggregated indicators (e.g., proportion of students with low attention, average fatigue level) as trend curves or compact heatmaps. This allows teachers to react during class to sudden drops in engagement by adjusting pacing, questioning strategies, or instructional modalities, and to reflect after class on when and where engagement declined to inform lesson redesign and targeted support. Over longer periods, aggregated statistics across lessons help administrators identify structural inefficiencies (such as overloaded time slots or problematic subject scheduling) and optimize timetables and resource allocation. From an educational economics perspective, this workflow converts previously unobservable behavioral processes into measurable process data, enabling more efficient use of limited teaching resources and supporting equity-oriented decision-making in resource-constrained schools, with modest hardware cost due to edge-based deployment.

4.6. Limitations and Future Work

Although the proposed study provides a comprehensive evaluation of multimodal student-state recognition and edge deployment, several limitations remain with respect to the core challenges of noisy sensing, multimodal fusion, and resource-constrained inference, and these should be addressed in future research. First, the dataset was collected from multiple rural primary schools and one county-level middle school and primarily

reflects typical classroom behaviors in this specific regional and cultural context. While the experiments demonstrate that LightEdu-Net is robust to sensor noise and heterogeneous modalities under these conditions, its generalization ability has not yet been fully validated for other subject types (e.g., laboratory sessions, project-based learning), special learning behaviors (e.g., students with learning difficulties), or classrooms with different cultural, linguistic, or curriculum backgrounds. In addition, the current dataset, though multimodal, still focuses on a limited set of physiological and environmental variables. Future work should expand data collection across more diverse schools and regions, include richer behavioral and contextual modalities, and explore unified annotation and preprocessing standards so that the noise-resilient and fusion mechanisms of LightEdu-Net can be systematically evaluated and adapted in broader settings. Second, although the proposed architecture achieves real-time inference on mainstream edge platforms such as Jetson Nano, Raspberry Pi 4B, and Intel NUC, the current design is primarily optimized for these mid-range devices. Under ultra-low-power or highly constrained hardware (e.g., wearable devices, low-end tablets, or legacy educational terminals), the computational cost of cross-modal attention and feature alignment remains a bottleneck, and the latency may no longer satisfy strict real-time requirements. The present SNAM, CMAF, and EAKD modules effectively alleviate the original deployment challenge within the tested resource envelope, but they do not yet provide a principled solution for the extreme edge regime. Future research should therefore investigate more efficient low-rank and sparse attention structures, event-driven or adaptive-inference strategies that selectively update only salient segments, and lighter distillation frameworks tailored to on-device training or continual learning. These directions will be important for further reducing computational and energy overhead while preserving the robustness and multimodal discriminative capability demonstrated in this work.

5. Conclusions

This study addresses a core challenge in deploying intelligent tutoring systems within resource-constrained educational environments, namely, how to perform reliable multimodal student-state recognition under sensor noise, heterogeneous inputs, and limited computational resources. To tackle this problem, we propose LightEdu-Net, a lightweight multimodal framework that combines a sensor-noise adaptive module to stabilize feature quality, a cross-modal attention fusion mechanism to enhance information interaction across visual, physiological, and environmental signals, and an edge-aware knowledge distillation strategy to preserve discriminative capacity on low-power devices. Experimental evaluations demonstrate clear progress in overcoming the identified challenges: LightEdu-Net achieves 92.4% accuracy with balanced precision and recall and maintains stable performance under noise perturbation while supporting real-time inference on Jetson Nano, Raspberry Pi 4B, and Intel NUC. These results confirm that the proposed solutions improve robustness, fusion expressiveness, and deployability, providing a viable pathway for multimodal perception in rural and under-resourced classrooms.

Despite these advances, several limitations remain when addressing the deployment challenge in its full scope. The current approach has been validated primarily on a limited set of schools and typical classroom settings, and its generalization to diverse learning activities, cultural contexts, or extreme hardware constraints has not been fully established. Moreover, although latency and power consumption are acceptable for mainstream edge devices, further reductions will be required for ultra-low-power platforms such as wearable systems or legacy terminals. Future work will therefore focus on expanding multimodal datasets across broader regions, investigating more efficient attention and distillation mechanisms, and exploring privacy-preserving training strategies such as federated learning

to enable scalable deployment across multiple schools. These directions will further enhance the practicality and sustainability of LightEdu-Net and accelerate its application in real-world educational environments.

Author Contributions: Conceptualization, C.H., Y.C., B.Z., X.C. and Y.Z.; Data curation, Z.Z. and J.Z.; Funding acquisition, Y.Z.; Methodology, C.H., Y.C., B.Z. and X.C.; Project administration, Y.Z.; Resources, Z.Z. and J.Z.; Software, C.H., Y.C., B.Z. and X.C.; Supervision, Y.Z.; Validation, J.Z.; Visualization, Z.Z.; Writing—original draft, C.H., Y.C., B.Z., X.C., Z.Z., J.Z. and Y.Z.; C.H., Y.C., B.Z. and X.C. contributed equally to this work. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China grant number 61202479.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kayumova, M.; Akbarova, S.; Azizova, M.; Hojiyeva, I.; Karimova, F.; Makhmudova, F. Data Driven Teaching and Real Time Decision Making in Education Management. In Proceedings of the 8th International Conference on Future Networks & Distributed Systems, Marrakech, Morocoo, 11–12 December 2024; pp. 756–763.
2. Trabelsi, Z.; Alnajjar, F.; Parambil, M.M.A.; Gochoo, M.; Ali, L. Real-time attention monitoring system for classroom: A deep learning approach for student's behavior recognition. *Big Data Cogn. Comput.* **2023**, *7*, 48. [CrossRef]
3. Wei, N.; Yang, F.; Muthu, B.; Shanthini, A. Human machine interaction-assisted smart educational system for rural children. *Comput. Electr. Eng.* **2022**, *99*, 107812. [CrossRef]
4. TS, A.; Guddeti, R.M.R. Automatic detection of students' affective states in classroom environment using hybrid convolutional neural networks. *Educ. Inf. Technol.* **2020**, *25*, 1387–1415.
5. Nye, B.D. Intelligent tutoring systems by and for the developing world: A review of trends and approaches for educational technology in a global context. *Int. J. Artif. Intell. Educ.* **2015**, *25*, 177–203. [CrossRef]
6. Hwang, K.A.; Yang, C.H. Automated inattention and fatigue detection system in distance education for elementary school students. *J. Educ. Technol. Soc.* **2009**, *12*, 22–35.
7. Pabba, C.; Kumar, P. An intelligent system for monitoring students' engagement in large classroom teaching through facial expression recognition. *Expert Syst.* **2022**, *39*, e12839. [CrossRef]
8. Alsayigh, H.K.S.; Khidhir, A.S.M. Using IoT to predict student attention levels in e-learning classes: A review. *AIP Conf. Proc.* **2024**, *3091*, 040009.
9. Gupta, S.; Kumar, P.; Tekchandani, R. A machine learning-based decision support system for temporal human cognitive state estimation during online education using wearable physiological monitoring devices. *Decis. Anal. J.* **2023**, *8*, 100280. [CrossRef]
10. Chen, S.; Hu, Z.; Li, S.; Hu, X.; Liu, G.; Wen, W. Autonomic nervous pattern recognition of students' learning states in real classroom situation. *IEEE Trans. Comput. Soc. Syst.* **2021**, *10*, 220–233. [CrossRef]
11. Miao, Y. Online and offline mixed intelligent teaching assistant mode of english based on mobile information system. *Mob. Inf. Syst.* **2021**, *2021*, 7074629. [CrossRef]
12. Sajja, R.; Sermet, Y.; Cikmaz, M.; Cwiertny, D.; Demir, I. Artificial intelligence-enabled intelligent assistant for personalized and adaptive learning in higher education. *Information* **2024**, *15*, 596. [CrossRef]
13. Liu, Y.; Chen, L.; Yao, Z. The application of artificial intelligence assistant to deep learning in teachers' teaching and students' learning processes. *Front. Psychol.* **2022**, *13*, 929175. [CrossRef]
14. Yang, Y. Design and Implementation of Algorithm-Based Intelligent Auxiliary System for Physical Education Teaching Equipment. *Int. J. High Speed Electron. Syst.* **2025**, *8*, 2540571. [CrossRef]
15. Wang, J. Generative Pre-Trained Transformer-4 based on English Assistant Teaching System in Higher Vocational Colleges. In Proceedings of the 2024 First International Conference on Software, Systems and Information Technology (SSITCON), Tumkur, India, 18–19 October 2024; IEEE: New York, NY, USA, 2024; pp. 1–5.
16. Wang, J.; Lu, T.; Li, L.; Huang, D. Enhancing personalized search with AI: A hybrid approach integrating deep learning and cloud computing. *J. Adv. Comput. Syst.* **2024**, *4*, 1–13. [CrossRef]
17. Ji, X.; Sun, L.; Huang, K. The construction and implementation direction of personalized learning model based on multimodal data fusion in the context of intelligent education. *Cogn. Syst. Res.* **2025**, *92*, 101379. [CrossRef]

18. Zhang, Y.; He, S.; Wa, S.; Zong, Z.; Lin, J.; Fan, D.; Fu, J.; Lv, C. Symmetry GAN detection network: An automatic one-stage high-accuracy detection network for various types of lesions on CT images. *Symmetry* **2022**, *14*, 234. [CrossRef]
19. Rashid, H.A.; Ovi, P.R.; Busart, C.; Gangopadhyay, A.; Mohsenin, T. Tinym2net: A flexible system algorithm co-designed multimodal learning framework for tiny devices. *arXiv* **2022**, arXiv:2202.04303.
20. Jiang, D.; Shen, Z.; Zheng, Q.; Zhang, T.; Xiang, W.; Jin, J. Farm-LightSeek: An Edge-centric Multimodal Agricultural IoT Data Analytics Framework with Lightweight LLMs. *IEEE Internet Things Mag.* **2025**, *26*, 9982–10001. [CrossRef]
21. Xiong, Y.; Guo, X.; Xu, J. CNN-Transformer: A deep learning method for automatically identifying learning engagement. *Educ. Inf. Technol.* **2024**, *29*, 9989–10008. [CrossRef]
22. Cruz, Y.J.; Villalonga, A.; Castaño, F.; Rivas, M.; Haber, R.E. Automated machine learning methodology for optimizing production processes in small and medium-sized enterprises. *Oper. Res. Perspect.* **2024**, *12*, 100308. [CrossRef]
23. Pei, Y.; Lu, G. Design of an intelligent educational evaluation system using deep learning. *IEEE Access* **2023**, *11*, 29790–29799. [CrossRef]
24. Bouafoud, C.; Zine-Dine, K.; Madani, A. The Evolution of Transformers in Education: A Literature Review. In Proceedings of the 2024 International Conference on Circuit, Systems and Communication (ICCSC), Fez, Morocco, 28–29 June 2024; IEEE: New York, NY, USA, 2024; pp. 1–7.
25. Sun, P.; Gu, L. Fuzzy knowledge graph system for artificial intelligence-based smart education. *J. Intell. Fuzzy Syst.* **2021**, *40*, 2929–2940. [CrossRef]
26. Dan, Y.; Lei, Z.; Gu, Y.; Li, Y.; Yin, J.; Lin, J.; Ye, L.; Tie, Z.; Zhou, Y.; Wang, Y.; et al. Educhat: A large-scale language model-based chatbot system for intelligent education. *arXiv* **2023**, arXiv:2308.02773.
27. Xu, Q.; Wu, Y.; Zheng, H.; Yan, H.; Wu, H.; Qian, Y.; Wu, Y.; Liu, B. Standardization in artificial general intelligence model for education. *Comput. Stand. Interfaces* **2025**, *94*, 104006. [CrossRef]
28. Gan, W.; Dao, M.S.; Zettsu, K.; Sun, Y. IoT-based multimodal analysis for smart education: Current status, challenges and opportunities. In Proceedings of the 3rd ACM Workshop on Intelligent Cross-Data Analysis and Retrieval, Newark, NJ, USA, 27–30 June 2022; pp. 32–40.
29. Li, C.; Liu, C.; Ju, W.; Zhong, Y.; Li, Y. Prediction of teaching quality in the context of smart education: Application of multimodal data fusion and complex network topology structure. *Discov. Artif. Intell.* **2025**, *5*, 19. [CrossRef]
30. Chango, W.; Lara, J.A.; Cerezo, R.; Romero, C. A review on data fusion in multimodal learning analytics and educational data mining. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2022**, *12*, e1458. [CrossRef]
31. Luo, Z.; Zheng, C.; Gong, J.; Chen, S.; Luo, Y.; Yi, Y. 3DLIM: Intelligent analysis of students' learning interest by using multimodal fusion technology. *Educ. Inf. Technol.* **2023**, *28*, 7975–7995. [CrossRef]
32. Shi, W.; Nie, Z.; Shi, Y. Application and Challenges of Multi-Modal Data Fusion in Adaptive Learning System. In Proceedings of the 2024 6th International Conference on Computer Science and Technologies in Education (CSTE), Xi'an, China, 19–21 April 2024; IEEE: New York, NY, USA, 2024; pp. 82–87.
33. Qianyi, Z.; Zhiqiang, L. Research on multimodal based learning evaluation method in smart classroom. *Learn. Motiv.* **2023**, *84*, 101943. [CrossRef]
34. Zhao, F.; Zhang, C.; Geng, B. Deep multimodal data fusion. *ACM Comput. Surv.* **2024**, *56*, 216. [CrossRef]
35. Dai, Z.; Zhang, Q.; Zhao, L.; Zhu, X.; Zhou, D. Cloud-edge computing technology-based internet of things system for smart classroom environment. *Int. J. Emerg. Technol. Learn. (iJET)* **2023**, *18*, 79–96. [CrossRef]
36. Kumari, K.A.; Sadasivam, G.S.; Dharani, D.; Niranjanamurthy, M. *Edge Computing: Fundamentals, Advances and Applications*; CRC Press: Boca Raton, FL, USA, 2021.
37. Shuvo, M.M.H.; Islam, S.K.; Cheng, J.; Morshed, B.I. Efficient acceleration of deep learning inference on resource-constrained edge devices: A review. *Proc. IEEE* **2022**, *111*, 42–91. [CrossRef]
38. Bauravindah, A.; Fudholi, D.H. Lightweight models for real-time steganalysis: A Comparison of MobileNet, ShuffleNet, and EfficientNet. *J. RESTI (Rekayasa Sist. Dan Teknol. Inf.)* **2024**, *8*, 737–747. [CrossRef]
39. Ahmadi, M.; Mirmahboub, B.; Karimi, N.; Khadivi, P.; Samavi, S. AI-Assisted Breast Cancer Classification: A Deep Learning Model Integrating MobileNet and ShuffleNet Features. In Proceedings of the 2025 IEEE World AI IoT Congress (AIIoT), Seattle, WA, USA, 28–30 May 2025; IEEE: New York, NY, USA, 2025; pp. 280–284.
40. Khediri, N.; Ben Ammar, M.; Kherallah, M. A real-time multimodal intelligent tutoring emotion recognition system (MITERS). *Multimed. Tools Appl.* **2024**, *83*, 57759–57783. [CrossRef]
41. Xia, C. Research on the Application of Large Language Models in Intelligent Tutoring System. In Proceedings of the 2025 3rd International Conference on Image, Algorithms, and Artificial Intelligence (ICIAAI 2025), Singapore, 23–25 May 2025; Atlantis Press: Dordrecht, The Netherlands, 2025; pp. 993–1003.
42. You, C.; Lu, H.; Li, P.; Zhao, X.; Yao, Y. AI-Driven Intelligent Learning Companions: A Multimodal Fusion Framework for Personalized Education. In Proceedings of the 2025 IEEE 34th Wireless and Optical Communications Conference (WOCC), Macau, China, 20–22 May 2025; IEEE: New York, NY, USA, 2025; pp. 424–428.

43. El Hajji, M.; Ait Baha, T.; Berka, A.; Ait Nacer, H.; El Aouifi, H.; Es-Saady, Y. An Architecture for Intelligent Tutoring in Virtual Reality: Integrating LLMs and Multimodal Interaction for Immersive Learning. *Information* **2025**, *16*, 556. [CrossRef]
44. Jia, J.; He, Y.; Le, H. A multimodal human-computer interaction system and its application in smart learning environments. In Proceedings of the International Conference on Blended Learning, Bangkok, Thailand, 21–24 July 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 3–14.
45. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
46. Howard, A.; Sandler, M.; Chu, G.; Chen, L.C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. Searching for mobilenetv3. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1314–1324.
47. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
48. Ab Wahab, M.N.; Nazir, A.; Ren, A.T.Z.; Noor, M.H.M.; Akbar, M.F.; Mohamed, A.S.A. Efficientnet-lite and hybrid CNN-KNN implementation for facial expression recognition on raspberry Pi. *IEEE Access* **2021**, *9*, 134065–134080. [CrossRef]
49. Barbieri, L.; Brambilla, M.; Stefanutti, M.; Romano, C.; De Carlo, N.; Roveri, M. A tiny transformer-based anomaly detection framework for IoT solutions. *IEEE Open J. Signal Process.* **2023**, *4*, 462–478. [CrossRef]
50. Çökmez, G.M.; Zhang, Y.; Schroers, C.; Aydin, T.O. CLIP-Fusion: A Spatio-Temporal Quality Metric for Frame Interpolation. In Proceedings of the 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Tucson, Arizona, 28 February–4 March 2025; IEEE: New York, NY, USA, 2025; pp. 7450–7459.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

sensors

MDPI

Article

Towards Explainable Multimodal Sensing for Swimming Analysis: Early Findings from the SWIM-360 Project

Vanessa Camilleri *, Reno Yuri Camilleri, Mark Fialovszky, Daniel Pace, Dylan Seychell and Matthew Montebello

Department of AI, University of Malta, MSD 2080 Msida, Malta; matthew.montebello@um.edu.mt (M.M.)
* Correspondence: vanessa.camilleri@um.edu.mt

Abstract

Swimming performance analysis increasingly depends on multimodal sensing systems that capture physiological and biomechanical signals in real-world aquatic environments. While progress has been made in sensor fidelity and automated analysis, the interpretability of these systems remains limited, constraining their uptake in coaching practice. This paper presents early findings from the SWIM-360 project, which investigates how explainable artificial intelligence (XAI) can support transparent and actionable insights for swimming performance. We report preliminary results from EO SwimBETTER and TrainRed sensors, together with proof-of-concept outputs from video-based pose estimation. In parallel, we introduce mock-up visualisations and interaction concepts designed to elicit coach feedback on requirements for explainability. A qualitative questionnaire with eight professional swimming coaches was conducted to elicit requirements for explainable feedback. Their responses informed the design of a multimodal, coach-centred explainability framework. Rather than providing a fully integrated model, the paper proposes a methodological framework that combines multimodal sensing with explainability-driven design principles. Our findings highlight both the feasibility and the challenges of translating sensor data into interpretable knowledge for athletes and coaches. By embedding explainability at the earliest design stage, this study proposes an explainability-driven design framework linking multimodal sensing and user requirements. These early findings highlight how XAI principles can guide the creation of trustworthy, coach-centred decision-support tools in aquatic sports.

Keywords: artificial intelligence (AI); explainable artificial intelligence (XAI); multimodal sensing; swimming analysis; wearable sensors; sensor technology; video-based pose estimation; human–computer interaction

1. Introduction

In recent years, sports science has made big steps forward through the use of wearable sensors, computer vision, and artificial intelligence (AI). Swimming is a good example, since performance depends on a mix of biomechanics, physiology, and technique. Tools such as inertial measurement units (IMUs) can measure stroke timing and turns, near-infrared spectroscopy (NIRS) can track muscle oxygen levels linked to fatigue, and video-based pose estimation can capture joint movement without markers. Together, these technologies provide much richer information than traditional stopwatches or coach observations.

The challenge is that the results from these systems are often difficult to understand. Many AI models act as "black boxes," producing outputs that are not easy for coaches or athletes to interpret. Some research has focused mainly on improving accuracy [1,2], while

others argue that clear explanations are essential for trust and real use in training [3,4]. Explainable AI (XAI) methods are beginning to address this gap. For example, SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations) can show which features of sensor data influenced a prediction, while visual overlays and attention maps can highlight which body parts or motion phases the model used to classify a stroke. Still, in swimming, these approaches remain relatively underexplored.

This paper presents early results from the SWIM-360 project, which explores how multimodal sensing and explainability can work together in swimming. Given the small participant pool, this study adopts an exploratory scope and is designed to collect qualitative insights rather than establish generalisable findings. We report findings from IMU and NIRS sensors , including data collected using TrainRed near-infrared spectroscopy sensors (Train. Red B.V., Gelderland, The Netherlands) to monitor muscle oxygen saturation (SmO2) and EO-SwimBETTER IMU-based sensors (EO, Sydney, Australia) for biomechanical motion capture, alongside proof-of-concept pose estimation outputs. We also share mock-up visualisations designed to gather coach feedback. Our aim is to test the feasibility of capturing and combining these data sources and to explore how explainability can turn them into insights that are useful in practice. By making explainability a central design goal from the start, this work contributes to building trustworthy, coach-friendly tools for swimming performance analysis. Overall, this study contributes a novel integration of explainability-driven design with multimodal sensing for swimming. Unlike prior work focusing on algorithmic accuracy, we foreground human interpretability by gathering empirical coach input to inform system development.

2. Background and Literature

2.1. Multimodal XAI for Sensor-Video Fusion

Multimodal explainability (MXAI) techniques aim to clarify how each data modality (e.g., wearable sensors, video, physiological signals) contributes to a model's decisions. A common approach is to disentangle the model's prediction into modality-specific contributions, plus any interaction effects between modalities [3]. For example, DIME (Disentangled Multimodal Explanations) extends LIME to multimodal models by applying local feature attribution on each modality separately and on their fused representation [3]. This yields an importance score per modality, helping quantify, say, how much an IMU sensor stream versus a video stream influenced a swimming stroke classification. Other methods incorporate explainability into the fusion mechanism itself. Attention-based fusion layers can produce learned weights that indicate each modality's influence on the output [3]. In a sentiment analysis study, researchers computed a numerical "effectiveness metric" from fusion parameters to show how modality interactions (e.g., audio vs. text) contributed to predictions [3]. Likewise, capsule networks have been used to estimate the dynamic contribution of each modality over time [3]. As a simple alternative, ablation analysis can directly score modality importance by measuring performance drop when each modality is withheld [3]. For instance, removing the video input might significantly degrade a model's accuracy if vision cues were critical, indicating the video's high importance. Together, these MXAI techniques provide a fusion-level explanation that assigns credit (or blame) to each modality for a given outcome, which is invaluable in a swimming context for understanding whether performance predictions are driven more by biomechanical (video) cues, physiological fatigue signals, or sensor-measured motion patterns.

2.2. Hierarchical Prototype Explanations (HIPEs) for Video Actions

Hierarchical interpretability frameworks explain predictions at multiple semantic levels of detail. HIPE (Hierarchical Prototype Explainer) is a recent approach that learns

case-based prototypes organised in a hierarchy to interpret video action recognition [1]. Instead of providing a single flat explanation, HIPE aligns with human-like description: for example, an activity might be explained as a water sport, then specifically as swimming, and finally as the butterfly stroke. Gulshad et al. introduce HIPE to connect lower-level class prototypes to higher-level parent classes, leveraging hyperbolic embeddings to naturally encode concept hierarchies [1]. The model identifies prototypical video snippets for fine-grained classes (e.g., a prototype clip of butterfly arm motion) as well as prototypes for broader parent categories (e.g., generic swimming motions). At inference time, the prediction can be explained through a multi-level chain: "the model predicts this is a butterfly stroke because it looks like the prototype for a butterfly, which in turn is a kind of swimming". This yields a human-understandable rationale like "swimming → butterfly → specific stroke technique."

The hierarchical structure also helps in cases of uncertainty: if the model isn't confident at the specific stroke level, it can fallback to explaining the prediction at the parent level (e.g., simply as "water sport") [1]. An important finding is that incorporating hierarchy can improve the accuracy-interpretability trade-off; HIPE maintained competitive accuracy on the UCF-101 video action dataset while providing richer explanations, outperforming a non-hierarchical prototype model [1]. For sports analytics, this approach is promising—a classification of a swimmer's activity could be transparently broken down into levels (sport, stroke, sub-technique), aiding coaches in understanding not just what the model predicts, but why at multiple levels of granularity. One implementation challenge is that a predefined taxonomy of actions is required, and the model must learn meaningful prototypes for each category level, which can be complex for subtle skill variations in swimming.

2.3. Motion-Relevance Saliency in Video Sequences

For video-based analysis, a key explainability question is: which parts of the motion were most relevant to the prediction? Saliency maps and temporal attention techniques address this by highlighting the frames or spatial regions that the model focused on. A direct extension of image saliency to video is to apply methods like Grad-CAM on each frame or on 3D convolutional blocks. Saha et al. revisited Grad-CAM for action recognition and showed that it can indeed pinpoint motion-relevant regions: for a tennis swing video, Grad-CAM correctly highlighted the moving tennis racket as the salient region leading to the "swing" classification [5]. This indicates the model latched onto the motion of the racket rather than the background pixels. However, a naive frame-by-frame application of saliency ignores temporal context [5]. A significant challenge is capturing temporal saliency—which frames (or sub-sequences) were most crucial [5]. Grad-CAM applied over a sequence must ensure the order of frames is accounted for; if a video's frames are shuffled or played in reverse, the explanation should reflect changed causality [5]. Research has found that simply treating each frame independently can be misleading, since action understanding depends on the sequence of movements [5].

To tackle this, some explainability methods explicitly use motion information. For example, Zhi et al. propose a motion-guided frame sampling strategy (MGSampler) that inherently emphasises high-motion segments: they compute frame-to-frame differences and assume frames with large changes carry more information about the action [6]. By grouping video frames by motion magnitude, their sampler "explains" its selection by favouring segments where the important action occurs (e.g., the frames during a swimmer's arm pull underwater, not the gliding between strokes) [6]. In general, temporal saliency maps can be created by attributing importance scores to each frame or time window, often visualised as a timeline highlighting peaks where the model paid attention. Such motion-relevance explanations have been applied in few-shot action recognition and video question-answering

to identify key moments of change or interaction. The main implementation hurdle is ensuring spatiotemporal consistency: the highlighted regions must move with the athlete and persist over the critical frames, requiring models to integrate spatial and temporal attention. Despite these challenges, temporal saliency is vital in sports—it could show a coach which part of a dive or turn caused a judging score prediction or which stroke in a lap triggered a fatigue warning, thereby making the AI's decision process more transparent.

2.4. Attention-Based Fusion for Wearable Sensors

When multiple wearable sensors (e.g., IMUs on wrists, trunk, legs) are used to classify athletic movements, attention mechanisms can provide built-in explanations of the data fusion. Instead of treating all sensor inputs equally, an attention-based model learns to assign a weight to each sensor modality or even each time step, indicating its relevance to the current prediction. For example, Mahmud et al. introduced a self-attention network for human activity recognition (HAR) that includes a sensor modality attention layer [7]. This layer produces a set of attention weights (one per sensor stream) that essentially rank the importance of each sensor's signal in contributing to the recognised activity [7]. In their experiments, these learned weights formed sensor-level attention maps that were "intuitively explainable", allowing the authors to confirm that, say, for a "walking" activity, the foot-mounted IMU was given higher weight than hand-mounted ones [7]. Earlier, Ma et al. (AttnSense) [8] had combined CNN and GRU layers with both temporal attention and sensor-level attention, enabling the model to focus on salient time periods and sensor channels for multimodal HAR [7–9]. Such attention-based fusion not only improved accuracy but also yielded interpretable insights: the model could identify which sensor was most informative for a given action (e.g., a gyroscope on the dominant arm is key for detecting a butterfly stroke, while an accelerometer on the head might be less relevant). A strength of attention weights is that they are directly readable from the trained model for any given input, offering a form of local explanation (per prediction) as well as a global explanation (average attention over many samples) about sensor importance.

Attention maps have been used to build simple visualisations—for instance, a bar chart highlighting that "Sensor 1 (wrist IMU) contributed 50% to the detection of 'flip turn', whereas Sensor 2 (ankle IMU) contributed only 20%". However, it is important to note that raw attention weights should be interpreted with caution. Recent studies have debated whether attention weights truly align with causal importance [3]. In some cases, attention-based explanations can conflict with gradient-based attributions, meaning high weight doesn't always imply a large effect on the output [3]. Despite this caveat, in practice, attention fusion has proven very useful in wearable sports analytics because it naturally integrates into deep models and provides sensor-level interpretability without separate post-hoc analysis. Implementation challenges include ensuring the attention mechanism doesn't just attend to sensors with higher noise or magnitude, and validating that what the attention highlights makes sense to human experts (e.g., does the model focus on the IMU that a coach would intuitively consider most important for a given skill?). When designed and validated carefully, attention-based sensor fusion can greatly enhance trust in multimodal sport AI systems by clearly identifying which streams of data are driving each decision.

2.5. AI Explainability 360 Toolkit and Taxonomy

To navigate the broad landscape of XAI techniques, researchers at IBM proposed the AI Explainability 360 (AIX360) toolkit, which includes a taxonomy of explainability methods [4,10]. The AIX360 taxonomy is an attempt to systematically organise methods by criteria such as intrinsic vs. post-hoc, global vs. local, and the target stakeholder of the

explanation [4]. The premise is that no single explainability method suits all needs—for example, a coach, an athlete, and a data scientist might all require different explanation formats from the same system [4]. AIX360 provides a structured way to pick the right tool for the job. It classifies and implements diverse algorithms covering tabular data, time-series, text, and image modalities in one open-source library [4,10]. This is highly relevant for sports and biomechanics, where datasets are often multimodal (e.g., numerical performance metrics, video, and sensor streams together). Using the AIX360 framework, one can mix and match explainer methods for different components of a system. For instance, in a multimodal injury prediction model with both questionnaire data and motion capture video, the taxonomy might suggest using a feature attribution explainer (like SHAP) for the tabular questionnaire features, and a visual attribution (like Grad-CAM) for the video component [3]. Tjoa et al. [2] demonstrated a combined approach in a medical context: they applied SHAP to handle patient metadata (tabular features) while using Grad-CAM on imaging data, creating a joint explanation for a skin lesion diagnosis model [2].

A similar approach could be taken in a swimming analytics system: SHAP values could explain sensor-derived features (e.g., highlighting that "muscle O_2 saturation drop was a strong contributor to the fatigue flag"), while visual saliency overlays explain video-based features (e.g., highlighting the moment a swimmer's elbow dropped, causing a technique fault). The AIX360 toolkit not only provides these algorithms but also metrics to evaluate explanations and tutorials to guide practitioners [4]. By following a taxonomy-driven approach, developers can ensure that for each modality and user type in a sports system, an appropriate explainability method is chosen. In summary, AIX360 encourages a systematic, multi-method explainability strategy: rather than relying on a single "one-size-fits-all" explainer, it promotes orchestrating several techniques (rules, feature attributions, prototypes, saliency maps, etc.) so that every aspect of a complex model (and every stakeholder) is covered by an understandable explanation [4]. This philosophy aligns well with the needs of high-performance sports environments, where coaches might demand human-interpretable rules and summaries, whereas sports scientists might appreciate quantitative feature impact charts—all of which can be integrated under a unified XAI framework.

2.6. Traditional XAI Methods in a Sports Context

Many traditional explainability methods from machine learning can be adapted to the sports analytics domain, often by customising them to sensor or video data. For sensor-based models, common model-agnostic tools like SHAP and LIME are popular for explaining predictions. SHAP assigns each feature a contribution value for a particular prediction, derived from game-theoretic principles. In athlete monitoring research, SHAP has been used to identify influential factors in performance outcomes—for example, one study on predicting young athletes' fitness found that SHAP analysis could clearly rank the importance of features such as training load, age, and VO_2max in determining fitness levels [11]. This gave coaches insight into why the model might rate one athlete's fitness higher than another (e.g., a higher training load had a positive SHAP value for the fitness score). LIME can likewise be applied by perturbing sensor inputs (like slightly altering stride length or stroke count features) to see how the prediction changes; however, LIME's use in multimodal settings can be tricky, which is why methods like DIME were developed to handle each modality separately [3].

Aside from these post-hoc methods, simpler feature importance metrics are often reported when using tree-based models or linear models on sports data. For instance, a random forest that predicts swim race time from biomechanical metrics can output an importance value for each input (e.g., showing that stroke rate is the top contributor, followed

by turn time, etc.). Partial dependence plots (PDPs) are another classic tool: they chart the model's predicted outcome as a function of one feature, averaging out others. In a swim analysis context, a PDP could reveal a non-linear relationship, such as "beyond 50 strokes per minute, further increases in stroke rate yield diminishing improvements in speed," which is valuable for training decisions. One recent study compared different neural network models on wearable HAR data and applied Layer-wise Relevance Propagation (LRP) to interpret them; the LRP results showed, for example, that gyroscope Y-axis data was consistently the most informative sensor input for distinguishing activities, whereas certain other channels contributed little [12]. This kind of insight—that "rotation about the Y-axis is key for this movement"—aligns with domain knowledge and validates the model's focus. Furthermore, in sports settings, there is still a role for transparent, rule-based explanations. Traditional expert systems and simple threshold rules have long been used by coaches, precisely because their logic is easy to follow. For instance, a rule-based algorithm might flag "possible fatigue" if a swimmer's stroke rate drops by >10% while heart rate is high. Such approaches are highly interpretable and deterministic, as each decision can be traced to a clear rule [13]. Despite the limitations of handling complex patterns, these rule-based methods provide strong transparency and have been integrated even in modern hybrid systems (sometimes alongside machine learning) to ensure certain outputs remain understandable [13]. In summary, for sensor modalities, the practitioner has a spectrum of XAI options—from advanced methods like SHAP that quantify each feature's contribution, to straightforward if-then rules or feature rankings—and combining these can yield both rigorous and coach-friendly explanations (e.g., "SHAP says lap time is the top factor, and indeed whenever lap time > X, our rule triggers an alert").

For video and pose data, traditional explainability often takes the form of visualisations. Saliency maps (heatmaps overlayed on video frames) derived from methods like Grad-CAM, RISE, or guided backpropagation can show which body regions or objects in the scene influence the model's decision [3]. In action recognition tasks, these saliency techniques have been extended to highlight not only spatial areas but also correspond to specific frames [5], effectively indicating "when" and "where" the model paid attention. In a swimming video, a saliency overlay might light up the swimmer's kicking legs during a freestyle sprint if the kick technique drives the classification of the stroke intensity. Another approach is to leverage the outputs of pose estimation directly: since pose models produce interpretable joint coordinates, skeletal stick figures can be drawn on the video to represent the swimmer's motion. Many studies that use pose data for feedback simply present the reprojected skeleton or keypoint trajectories on the video. While this is technically a form of data visualisation rather than an explanation of a black-box model, it serves an explainability purpose by linking the analysis to something the coach can see—for example, highlighting an asymmetry in left vs. right-hand trajectories through the water.

Additionally, researchers have proposed overlaying vector fields or force arrows on athlete videos to visualise computed forces or motion directions. In swimming, this could mean displaying arrows along the swimmer's hand path indicating propulsive force magnitude and direction, thus explaining why a particular stroke was labelled inefficient (e.g., the force vectors show a lot of lateral slippage instead of forward thrust). Attention-based methods also translate to visual explanations: if a transformer model outputs an attention map over video frames, this can be depicted as frame-wise highlighting or a timeline curve showing peaks—analogous to a coach's eye focusing on the critical moment of a turn. The common thread is that for visual data, explanations tend to be visual. This aligns with coach feedback: heatmaps on a replay, annotated videos, and overlayed analytics are often more intuitively understood than numbers or text [3]. The main challenge is ensuring these visual explanations truly correlate with model reasoning (avoiding random or misleading

highlights) and that they can be generated efficiently (ideally in real-time for feedback during training). Nonetheless, adapting saliency and interpretability tools to sports videos is a highly active area, bringing computer vision explainability techniques into practical use on the pool deck.

2.7. Comparative Analysis and Implementation Challenges

Each of the above frameworks offers a different balance of insight, complexity, and usability, and combining them effectively in a sports analytics system requires careful consideration. A fundamental trade-off is between intrinsic interpretability and model accuracy [1,2]. Techniques like prototype models (HIPE) or rule-based systems bake interpretability into their structure, but they may sacrifice some predictive performance or require simplified model architectures. Conversely, post-hoc explainer methods (SHAP, LIME, saliency maps) can be applied to any high-performance model, but their reliability and fidelity become concerns. Researchers have noted that adding interpretability often comes at a cost—either in accuracy or in additional computation—though clever design can mitigate this [1]. For instance, HIPE's use of hyperbolic embeddings maintained accuracy while providing multi-level explanations, partially resolving the typical accuracy-explainability tension [1].

Another challenge is evaluating and validating explanations, especially in multimodal settings. Unlike predictive accuracy, there is no universally agreed metric for explanation quality. It is difficult to say whether an attention heatmap or a SHAP summary is a "good" explanation without reference to human judgment. In fact, different XAI methods can disagree—an attention-based explanation might highlight different features than a gradient-based explanation for the same prediction [3]. This lack of consistency raises questions: which explanation should we trust? Efforts are underway to develop benchmarks and ground-truth datasets for explainability, but currently, the evaluation of XAI in sports is often subjective. Experts might need to verify that what the model deems important (e.g., a certain kinematic pattern) aligns with coaching knowledge. The absence of ground-truth visual explanations (e.g., labelled "important" video segments from coaches) and standardised evaluation protocols makes comparative assessment of these techniques difficult [3]. This is an active research area, and improving it will help identify which XAI methods truly aid decision-making in practice.

A related issue is the faithfulness vs. interpretability dilemma: Some explanations are simple and appealing (like "Sensor X is most important"), but they risk oversimplifying the model's logic. There is a danger of coaches drawing false confidence from an explanation that only tells part of the story. For example, a partial dependence plot might show a clear trend of "more distance per stroke is better," but the model might also be internally using other variables and interactions that the PDP doesn't convey. Thus, combining multiple explanation types can be beneficial, providing both a high-level intuitive rule and a more detailed feature attribution, yet this also means more information for the user to process.

Real-time constraints pose another implementation challenge in sports environments. Unlike offline analysis, a coaching feedback system may need to generate explanations on the fly (e.g., right after a swimmer finishes a lap). Techniques like SHAP and LIME can be computationally intensive, as they require many model evaluations per explanation [3]. Approximations or simplified surrogate models might be needed to deliver explanations quickly. On the other hand, attention weights or prototypes are obtained directly from the model's forward pass and can be more real-time friendly. System designers must choose explainer methods that fit within time and resource budgets, possibly using pre-computed explanations or caching to speed things up.

Perhaps the most crucial factor is human interpretability and user acceptance. The end goal of XAI in sports is not just to satisfy academic interest, but to genuinely assist coaches

and athletes. Explanations should be presented in forms that stakeholders find meaningful, which may vary. A coach with a limited data science background might prefer a simple visual or a natural-language summary (e.g., "The AI flags a drop in elbow height as the reason for speed loss in your last 50 m"). A sports scientist or biomechanist might appreciate a more quantitative explanation (e.g., a chart of feature importances or a comparison to prototypical elite performance data). Therefore, tailoring the explanation to the end-user's expertise and cognitive needs is essential [3]. Research underscores that one-size-fits-all explanations often fail [4]—non-experts can be overwhelmed or misled by technical details, whereas domain experts may require more depth and precision [3]. Achieving the right personalisation means involving users in the design (co-operative design of dashboards, user studies on interpretability) and possibly offering multi-tiered explanations (an overview first, with the ability to drill down for more detail).

In comparing the techniques, Model-specific vs. agnostic methods are a key distinction. Prototype learning, attention mechanisms, and hierarchical models require designing (or at least retraining) the model in a special way—they offer built-in explanations which are often more direct and stable (since they reflect the model's actual components), but at the cost of flexibility. In contrast, agnostic methods like SHAP, LIME, or LRP can explain any model after the fact; they are very flexible and can be swapped out as needed, yet one must trust that these approximations accurately reflect the true reasoning of the model [3]. Ensemble or hybrid strategies (as advocated by AIX360) try to get the best of both worlds by using intrinsic interpretability where feasible and supplementing with post-hoc explanations elsewhere.

Lastly, integration and visualisation of explanations in a user interface is a non-trivial practical challenge. A multimodal system might generate several pieces of explanation: heatmaps on video, graphs for sensor feature importance, textual descriptions, etc. Presenting this coherently in a "coach-facing dashboard" without overwhelming the user is difficult [14]. Early prototypes in the literature have used traffic-light indicators, simple icons, or concise textual hints on dashboards to convey model findings. The design should avoid cognitive overload and focus on actionable insight: the explanation should answer a coach's implicit questions like "what should I look at in this performance?" or "why did the AI flag this segment?". Achieving this requires interdisciplinary effort—combining the robust algorithms from XAI research with HCI (Human-Computer Interaction) principles and domain knowledge in sports.

Figure 1 illustrates the conceptual architecture of the SWIM-360 system, which integrates multimodal sensing, explainability mechanisms, and user-centred feedback. The process begins with data capture from wearable inertial and near-infrared spectroscopy (NIRS) sensors together with dual-perspective underwater video. These data streams pass through calibration, synchronisation, and feature-extraction pipelines before being combined within a planned multimodal analytics layer. A dedicated explainability layer implements a hybrid XAI strategy that merges SHAP-based feature attribution for sensor data with visual overlays and rule-based heuristics for video insights. The resulting feedback is aggregated into a coach-facing interface that provides interpretable performance summaries and personalised alerts.

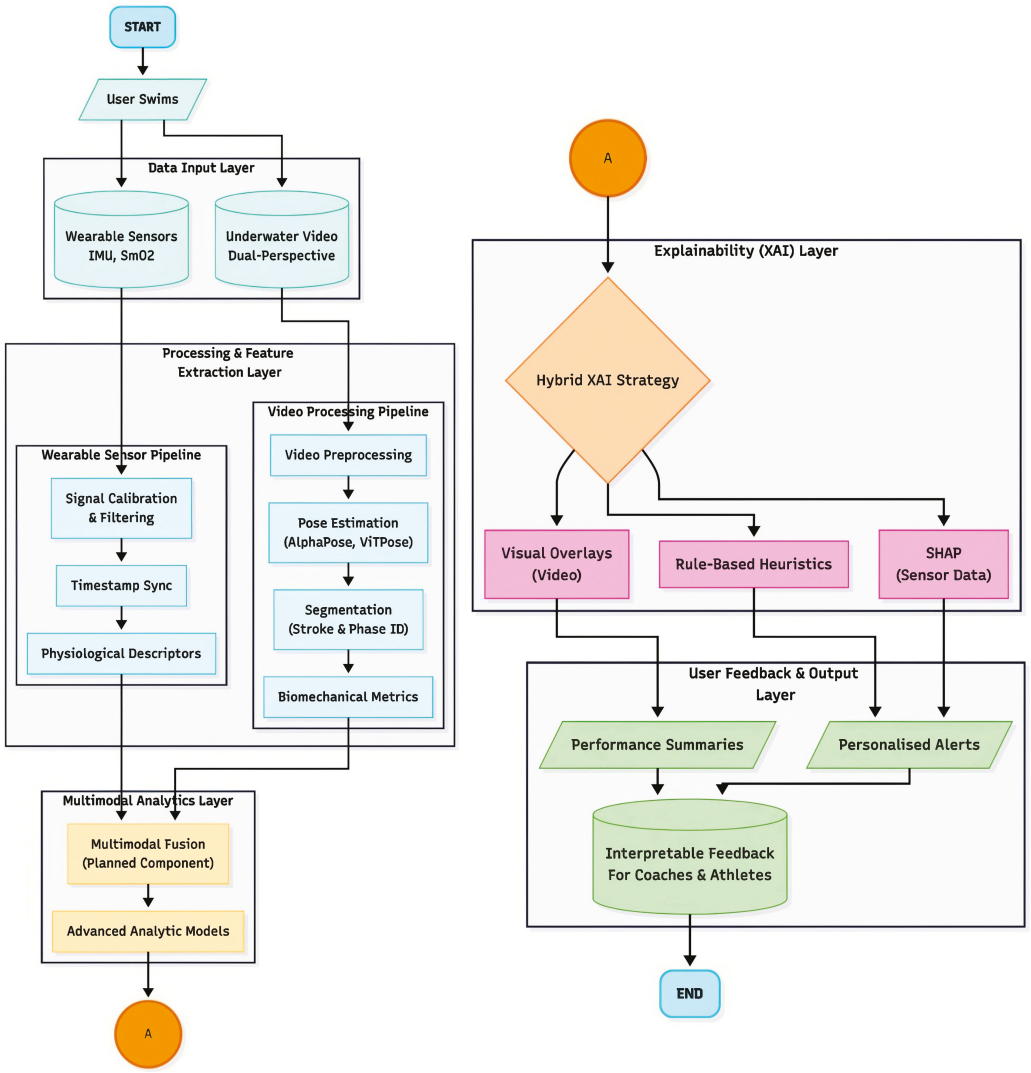


Figure 1. Conceptual architecture of the proposed SWIM-360 system. The workflow begins with data collection from wearable sensors (IMU, SmO2) and dual-perspective underwater video. The processing pipeline extracts physiological and biomechanical descriptors, aligns sensor and video data, and feeds them into a planned multimodal analytics layer. The Explainability (XAI) Layer applies a hybrid strategy, combining visual overlays, SHAP-based feature attributions, and rule-based heuristics to generate interpretable feedback for coaches and athletes through performance summaries and personalised alerts.

In summary, the reviewed literature highlights persistent gaps in multimodal explainability for sports analytics, including (1) limited integration of sensor and video modalities within a unified explainability framework, (2) the absence of domain-specific evaluation protocols for interpretability, and (3) minimal engagement of end-users, such as coaches, in the design and validation of AI systems.

To address these gaps, the SWIM-360 project applies explainability methods to real-world swimming analysis through an iterative, user-centred approach. Drawing on frameworks such as HIPE, attention-based fusion, and the AIX360 taxonomy, SWIM-360 translates abstract multimodal explainability principles into design requirements that are tested and refined in collaboration with coaches. By embedding explainability and usability considerations at every development stage, the project bridges the methodological divide between theoretical research and practical application in aquatic sport environments.

Feedback from pilot testing has guided refinements in multimodal data fusion, visual feedback, and usability design, ensuring that interpretability aligns with coaching practice

and training objectives. This iterative validation process addresses the lack of participatory design in prior studies and establishes a foundation for domain-specific evaluation of explainable outputs.

Collectively, these developments show how SWIM-360 operationalises multimodal explainability as both a technical and human-centred objective. Rather than benchmarking isolated models, the project focuses on transforming explainable AI methods into actionable, interpretable feedback mechanisms tailored to the needs of coaches. This approach advances the field by demonstrating how user-informed design can turn multimodal explainability research into transparent and trustworthy decision-support systems for performance analysis.

3. Methodology

3.1. Overall Review

The methodological approach adopted in this study followed an explainability-driven and human-centred design process, integrating principles from multimodal sensing, XAI, and participatory system design. The purpose was not only to test technical feasibility but to understand how coaches interpret, trust, and make use of multimodal feedback in swimming. To achieve this, a structured questionnaire was designed to capture professional coaches' expectations and preferences for explainable feedback, forming the empirical foundation for subsequent dashboard and model-development phases. Given the exploratory scope and small participant pool (n = 8), the study was designed to collect qualitative insights rather than produce statistically generalisable results. Participants were recruited through professional swimming coaching networks in Malta and the EU, with inclusion criteria requiring at least two years of coaching experience.

The questionnaire was conceived in direct response to the theoretical and technical insights reviewed in Section 2. As discussed there, the literature on explainable multimodal AI emphasises that successful implementation depends on aligning algorithmic transparency with the informational needs of human stakeholders [3,4]. Frameworks such as AIX360 [4] underline that no single explanation method suits all audiences; rather, the form and depth of explanation must adapt to the user's role and expertise. Coaches, who ultimately interpret sensor outputs in context, therefore became the focal participants in the early design phase of SWIM-360. The questionnaire served as a mechanism for eliciting these user requirements, so that later technical decisions could be informed by an evidence-based understanding of what kinds of explanations are meaningful in real coaching environments.

The design process began with a mapping of conceptual dimensions identified in Section 2, interpretability, modality importance, visual saliency, temporal explainability, and stakeholder relevance, to the structure of the questionnaire. The resulting questionnaire consisted of four sections (General Information, Visualisation Preferences, Practical Use, and Technique and Biomechanical Feedback), each targeting a distinct aspect of explainable system design. The format combined short multiple-choice and ranking items with open-ended responses to balance quantifiable comparison and qualitative insight. All items were developed from first principles derived from XAI theory and multimodal sensing research, ensuring that each question had a methodological justification rather than being merely descriptive.

The opening section established contextual variables relevant to explainability. Questions about the swimmers' level and the coach's country of practice acknowledged that information needs vary across competitive levels and cultural settings, echoing prior observations that explainability is context-dependent rather than universal [3]. Items on the number of swimmers per session helped define system scalability, reflecting parallels

with multimodal fusion models where attention mechanisms distribute importance across multiple input streams [7,8]. Respondents were also asked to indicate which feedback categories, techniques, physiology, performance, or injury prevention were most useful. These domains correspond to the sensor modalities explored earlier in the paper (IMU, NIRS, and video), linking the human-facing survey directly to the technical architecture of the SWIM-360 system. The first section concluded with a Likert-scale item evaluating the perceived importance of explainability itself, derived from AIX360's emphasis on stakeholder-specific interpretability needs [4]. This item served as a benchmark for assessing how strongly transparency is valued relative to raw performance accuracy in real coaching contexts.

The second section examined preferences for visual explanations, integrating lessons from research on perceptive and multimodal interpretability [2,5]. Since visual overlays such as saliency maps and motion highlights have been shown to enhance understanding in action-recognition models [5,6], respondents were asked to select their preferred visualisation type, video replay with highlights, colour-coded feedback, trend graphs, or side-by-side comparisons, and to rate how helpful each would be in everyday coaching. These questions directly operationalised the principle that explanations should exploit the modality through which the underlying data are experienced: in this case, the predominantly visual nature of swimming technique. By contrasting alternative display types, the survey also allowed investigation of how different representational forms support or hinder interpretability across varying levels of analytical detail.

The third section focused on practical use and temporal context, drawing from the literature on human-centred explainability and system usability. Questions on preferred data format (numerical values versus interpreted statements) explored the balance between raw data transparency and semantic summarisation, reflecting the ongoing discussion between intrinsic and post-hoc explanations [3]. Items about information density and timing of feedback (real-time versus post-session) addressed cognitive and operational constraints. Real-time explanation poses known computational challenges; SHAP and LIME, for example, require repeated model evaluation, whereas attention-based mechanisms offer faster, built-in interpretability [7,9]. Capturing user priorities here helped determine which technical strategies would be viable for future dashboard deployment. Finally, questions on device type and usage frequency grounded the design in real coaching environments, ensuring that the proposed system adapts not only to users' expertise but also to their working context, whether poolside on a tablet or during detailed desktop analysis.

The final section of the questionnaire concentrated on technique-specific and biomechanical explainability, linking directly to the sensor analysis components of SWIM-360. Respondents rated the importance of metrics such as body alignment, kick symmetry, breathing timing, and turn efficiency, features corresponding to those emphasised in prototype-based and attention-based models for HAR [1,8,12]. These items translated theoretical variables into domain-specific terms familiar to coaches, ensuring conceptual continuity between algorithmic interpretability and biomechanical reasoning. The open questions that followed invited participants to describe the key elements they would most want monitored, the challenges of performance analysis without technology, and the attributes that would make the SWIM-360 dashboard most useful. These qualitative prompts were guided by participatory design principles and the co-creation ethos advocated in explainable AI research [3]. They sought to uncover coaches' implicit mental models, the heuristics and visual cues they already rely on, so that future system explanations could align with established professional reasoning rather than introducing alien concepts.

The questionnaire underwent internal pilot testing with researchers familiar with aquatic performance analysis. Revisions improved clarity and reduced redundancy be-

tween overlapping items, while maintaining a balance between brevity and conceptual coverage. A mixture of scaled and open responses ensured both statistical and interpretive richness. Ethical procedures were integrated throughout the design. Open-ended responses were analysed thematically by two independent reviewers using an inductive coding approach. Agreement on key themes exceeded 85%, ensuring consistency of interpretation. No personal identifiers were collected; all participation was voluntary. The consent statement at the beginning of the questionnaire functioned as both an ethical safeguard and a methodological parallel to explainability principles, explicitly informing participants how their data would be used, mirroring the system's eventual obligation to explain its outputs to end-users. For a more detailed view on each question's format and decision, refer to Appendix A. The study was approved by the University of Malta Research Ethics Committee (Ref: UM-REC-2025-03)

3.2. Preliminary Results

The preliminary multimodal measurements confirm that the system can reliably acquire and synchronise heterogeneous data streams under real swimming conditions. The video-based pose-estimation component yielded stable joint detections throughout the underwater sequence, enabling the identification of stroke phases as seen in Figure 2, and key temporal events. These outputs support lap-to-lap comparisons of timing characteristics and stroke counts, providing a consistent basis for analysing stroke patterns as well as their temporal evolution. Although the video-processing module is still being tuned, the current results indicate that the overall approach is technically feasible and promising for future large-scale deployment.

Results	
Total Video Duration	32.02 seconds (0.53 minutes)
Total Strokes	31
Left Strokes	16
Right Strokes	15
Balance Ratio	0.032 (lower - more balanced)
Stroke Rate	58.1 strokes/minute
Stroke Timing Analysis	
Average Stroke Interval	126.4 frames (1.05 seconds)
Stroke Interval Variability	34.3 frames
Most Consistent Stroke Rate	56.9 strokes/minute

Figure 2. Summary of stroke-related metrics extracted from the video-based analysis. The table reports total stroke count, left–right stroke distribution, balance ratio, and overall stroke rate, followed by stroke-timing indicators derived from frame-level detections. These measures provide an initial quantitative overview of temporal stroke characteristics across the analysed segment.

The EO SwimBETTER force sensors (eo, Sydney, Australia) yielded detailed biomechanical profiles of hand-applied propulsion as seen in Figure 3. Early recordings revealed clear left–right disparities in both force magnitude and directional distribution. Such asymmetries may arise from imperfect stroke strategy, habitual compensatory patterns, or stroke-side breathing in freestyle, which naturally introduces rotational and timing differences. The resulting force-field characteristics appear sufficiently sensitive to detect subtle deviations in technique that might not be captured through visual inspection alone. Observing the persistence or lap-specific variation of these asymmetries may also offer valuable context for understanding mechanical consistency and overall stroke effectiveness.

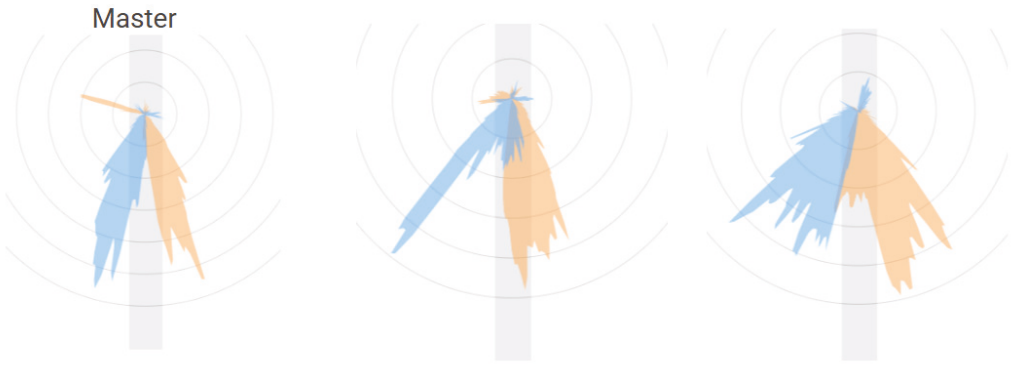


Figure 3. Comparison of force-field patterns derived from EO hand-mounted sensors for swimmers of different proficiency levels during freestyle. The reference "Master" swimmer (**left**) displays a compact and symmetric distribution of propulsive force vectors, whereas less experienced swimmers (**middle**,**right**) exhibit broader and more asymmetric patterns, indicating variations in hand trajectory, force orientation, and overall stroke consistency.

As seen in Figure 4, the TrainRed (Train. Red B.V., Gelderland, The Netherlands) NIRS measurements provided an additional physiological dimension by capturing the swimmer's muscle oxygenation dynamics. The SmO2 and HbDiff trends reflect the athlete's initial readiness at the start of the session and the subsequent evolution of local muscular demand during repeated stroke cycles. When interpreted alongside the mechanical and temporal indicators, these physiological profiles offer a broader understanding of the swimmer's condition and the factors that may influence variations in propulsive effectiveness. This alignment across modalities illustrates the potential of integrating metabolic-state information with biomechanical and temporal markers within a unified monitoring framework.

Taken together, these preliminary findings show that video-derived kinematics, hand-based propulsive force measurements and physiological signals can be examined within a unified multimodal system. Viewed together, the three modalities provide a clearer picture of stroke execution by relating temporal characteristics, such as stroke counts, phase timing and rhythm, to the measured propulsive forces and the observed physiological responses. This combined perspective helps describe how mechanical effectiveness, coordination and physiological state develop during the session, offering insights that are not available when each data stream is considered on its own.

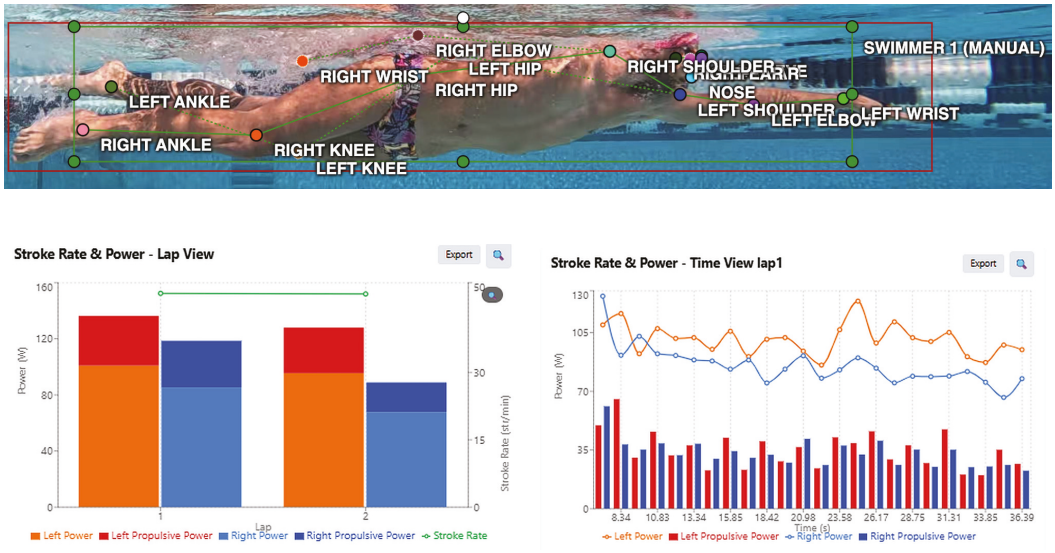
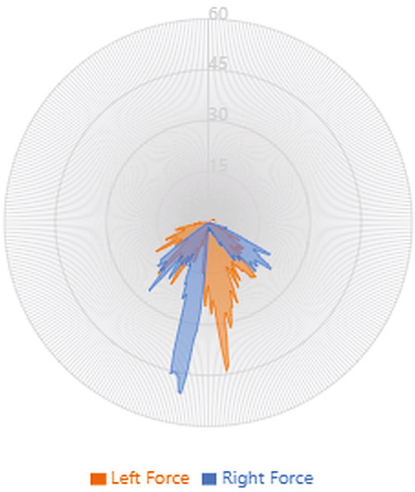


Figure 4. *Cont.*



Timestamp (seconds	Lap/Event		SmO2	HBDiff
0	0		63.18	17.5
0.07	0		63.23	17.51
0.22	0		63.32	17.54
0.27	0		63.44	17.57
0.36	0		63.6	17.62
0.52	0		63.79	17.67
0.56	0		64.01	17.73
0.67	0		64.25	17.8
0.76	0		64.5	17.87
0.85	0		64.75	17.94
1.03	0		65	18.01
1.06	0		65.26	18.08
1.19	0		65.51	18.15

Figure 4. Preliminary multimodal outputs combining video-based pose estimation, EO Swim-BETTER (eo, Sydney, Australia) force measurements, and Train.Red NIRS data. The top panel illustrates detected joint keypoints during an underwater stroke cycle. The middle plots display stroke-specific force and power characteristics, including left–right propulsive contributions and force-vector distributions. The lower table presents corresponding SmO2 and HbDiff values sampled throughout the lap, providing a concurrent physiological profile of muscular oxygenation and haemodynamic response.

4. Results

A total of eight responses were collected from practising swimming coaches who completed the SWIM-360 questionnaire. The purpose of this stage was exploratory: to capture qualitative insights into coaches' priorities, preferences, and perceived challenges regarding explainable multimodal feedback, rather than to achieve quantitative generalisation. All respondents provided informed consent and completed the full survey.

Although most coaches worked primarily with age-group swimmers, several also coached national-level and masters athletes, reflecting a range of competitive contexts. The typical number of swimmers monitored per session varied widely, from 1–4 individuals in small-group technical training to 11–20 in larger team environments. This spread indicates that feedback tools must accommodate both detailed, athlete-specific visualisation and efficient overviews suitable for group sessions.

Across respondents, there was near-unanimous agreement that the most useful forms of system feedback would include stroke technique, physical condition, and performance metrics. Several also emphasised injury prevention as an important though secondary concern. When ranking key coaching priorities, stroke efficiency and fatigue or recovery consistently appeared at the top, followed by race performance, with injury prevention often rated as somewhat less critical in day-to-day training contexts.

The perceived importance of explainability was high across all submissions. On the five-point scale, most coaches rated it either 4 or 5, indicating that they value systems capable of clarifying why particular results or alerts occur. The same tendency appeared when the question was repeated in the context of visual outputs, confirming that interpretability is not only valued conceptually but also expected to be visible in practical feedback displays.

When asked about preferred visualisation formats, the majority selected video replay with highlights, often accompanied by comments describing how athletes "can't see themselves swim" and how visual playback aids correction. Regarding overall interface preferences, seven of the eight coaches selected a customisable dashboard view as the screen they would most likely use daily. This suggests a strong demand for flexibility in how information is displayed, allowing users to tailor the interface to the focus of each training session. Only one respondent preferred a fixed "stroke technique view," indicating that while detailed biomechanical data are valued, they should be presented within a configurable structure.

When rating the usefulness of individual visual features, video highlights and colour-coded feedback received the highest ratings overall, typically 4–5 on a five-point scale. Graphs and trend lines were seen as helpful for tracking progress over time, but not necessarily as primary tools during live coaching. Side-by-side comparisons were considered particularly valuable for illustrating technique differences, especially in age-group and masters coaching, where reference to elite form can clarify corrective goals.

Half of the coaches preferred a side-by-side display of numerical data and system-generated interpretations (e.g., stroke rate alongside "below optimal range"). A smaller group preferred to see only the numerical outputs, while just one coach was satisfied with viewing only the interpreted results. This pattern suggests that presenting both forms of data together best supports understanding and trust in the system's feedback.

When asked about information density, nearly all respondents opted for key highlights with the option for details, underscoring a shared concern about information overload. This pattern indicates that an effective interface should prioritise clarity and brevity while still allowing expert users to explore deeper analytics when needed.

In terms of the timing of feedback, several respondents described real-time information as essential, while others considered post-session review sufficient. Those coaching smaller groups were more inclined to request immediate feedback, often citing the desire to "correct mistakes exactly when they happen," whereas those managing larger squads prioritised efficient summary reports.

Device preferences reflected the mobility of the poolside environment. Tablets and smartphones were the most frequently selected platforms, with some also mentioning laptop use for detailed analysis and large screens for group feedback. This distribution reinforces the need for a responsive design adaptable to multiple contexts of use. Regarding frequency, coaches indicated they would access the dashboard anywhere from every session to a few times per week, showing that the system must be both reliable for daily operation and lightweight enough to integrate smoothly into existing routines.

In the section on technique and biomechanical feedback, all metrics were rated as highly important, with mean scores typically between 4 and 5 on the importance scale.

The most valued parameters across strokes were arm pull depth and efficiency, body alignment and rotation, and start, turn, and push-off efficiency. Breathing timing and kick symmetry were also rated positively but were occasionally ranked slightly lower, suggesting that these may be more stroke-specific priorities.

Open-ended responses offered further qualitative insight. When asked to identify three key elements to monitor, coaches most often mentioned stroke rate, body position or alignment, underwater efficiency, and recovery metrics. One respondent highlighted "heart rate, strokes per minute, and comparisons with previous sessions," reflecting a desire to combine biomechanical and physiological feedback.

When describing the biggest challenges faced without technology, several themes emerged. Coaches frequently noted the difficulty of observing all swimmers simultaneously, the impossibility of athletes seeing themselves swim, and the risk of forgetting to provide timely corrections when working with larger groups. The absence of objective, quantitative data was also cited as a limitation, particularly concerning recovery and training load management.

Lastly, in response to what would make the SWIM-360 system most useful, coaches converged on the need for ease of use, mobile accessibility, and integration of real-time and comparative feedback. Some emphasised the importance of being able to share information directly with athletes to support self-reflection, while others requested features that could "make the efficiency improvement process shorter" by automatically highlighting performance deviations. For a more detailed view on all responses, refer to Appendix B.

While the sample size was limited (n = 8), the consistency across responses indicates reliable trends in coach preferences. The exploratory design is therefore appropriate for defining system requirements rather than statistical generalisation. A summary of the results is available in Table 1 below:

Table 1. Summary of Results.

Parameter	Range/Mode	Interpretation
Number of swimmers per session	1–20 (mean $\approx$ 10)	System must scale for individual and group use
Importance of explainability	Mean 4.6/5	High priority
Preferred visualisation	75% Video replay with highlights	Visual explanation preferred
Preferred device	87% Tablet/Smartphone	Mobile deployment essential

5. Discussion

The results of the coach questionnaire provide early empirical evidence for how explainability can be meaningfully integrated into multimodal sensing systems for swimming. It is important to note that these findings are based on feedback from a small sample of eight professional coaches. While this number limits statistical generalisation, the qualitative design was intentionally chosen to capture exploratory insights into user expectations and interpretability requirements at an early stage of development. Consequently, the findings should be viewed as formative rather than confirmatory. Nonetheless, the consistency of qualitative themes across submissions reinforces key points identified in the literature on MXAI, attention-based fusion, and user-centred interpretability. Together, these insights form a foundation for translating technical XAI concepts into practical, coach-facing design principles.

An additional methodological limitation concerns the questionnaire design. Questions Q6 and Q7 in Appendix A were phrased in a nearly identical way ("How important is it that the system explains why a certain result or alert was generated?" and "How important is it that the system explains why a result or alert appears?"). This redundancy may have led

participants to interpret the items as referring to the same construct, potentially reinforcing or inflating the apparent importance of explainability in the results. However, this overlap does not affect the qualitative insights derived from open-ended responses.

Despite these limitations, the responses reveal a clear pattern in how coaches value transparency and interpretability. Almost all participants rated explainability as "important" or "very important," both in general terms and in visual contexts. This aligns closely with the premise of the AIX360 framework [4], which argues that transparency should be calibrated to stakeholder needs. For the SWIM-360 project, the end-user is neither a data scientist nor an athlete, but a coach operating in a time-constrained, information-rich environment. The demand for explanations, why a particular result or alert appeared, demonstrates that black-box outputs are insufficient in applied sport contexts. As AIX360 highlights, explainability is not merely a model property but a form of communication between algorithm and user, and these findings affirm that coaches expect such dialogue to be part of the system interface.

The strong preference for visual interpretability provides further validation of the multimodal XAI approaches reviewed in Section 2. Video replay with highlights was by far the most popular feedback modality, followed by colour-coded overlays and side-by-side comparisons. These preferences correspond directly with techniques such as Grad-CAM saliency mapping [5] and motion-guided frame selection [6], which emphasise where and when the model attends within a sequence. Coaches' comments that "athletes can't see themselves swim" and "video analysis helps correct mistakes" indicate that visual explanations are not only comprehensible but functionally embedded in existing coaching workflows. Such visualisations effectively serve as intrinsic explainers, bridging the gap between algorithmic reasoning and the perceptual experience of movement.

Interestingly, several respondents requested comparative or reference-based views (e.g., "side-by-side with elite swimmers"), echoing the structure of HIPE [1]. HIPE models organise interpretations at multiple semantic levels, such as "swimming" $\rightarrow$ "freestyle" $\rightarrow$ "specific stroke technique," providing a human-like hierarchy of understanding. Translating this idea into SWIM-360 would allow explanations that operate on multiple conceptual scales, for instance, identifying an overall deviation in stroke class while also pinpointing the precise biomechanical feature causing inefficiency. The coaches' emphasis on contextual benchmarking demonstrates a preference for relational explanations, where meaning arises through comparison rather than absolute metrics.

The finding that most participants favoured customisable dashboard views reinforces the argument that explainability must be adaptive to user goals [3]. In multimodal XAI, adaptivity often refers to the fusion stage, where attention mechanisms or disentangling explainers (like DIME) adjust modality weights dynamically, but the same principle applies at the human-computer interface. A customisable layout allows the user to surface the modality or explanation type most relevant at a given moment, reflecting a practical form of interactive interpretability. The strong demand for such flexibility parallels the attention-based fusion frameworks proposed for wearable sensors [7,8], where the system learns to focus on the most informative inputs. Here, the coach plays an analogous role: directing attention across feedback modalities depending on training objectives.

Equally significant is the balance between raw and interpreted data preferred by participants. Half of the coaches requested both numerical and semantic outputs side by side, illustrating the dual requirement of transparency and abstraction. This resonates with the multimodal taxonomy proposed by Rodis et al. [3], in which local quantitative explanations (e.g., SHAP values for sensor features) are complemented by higher-level verbal or visual summaries. For SWIM-360, this finding suggests that explanations should

not replace raw performance data but contextualise them, ensuring that interpretability does not obscure domain-relevant detail.

Responses concerning timing and cognitive load mirror broader tensions discussed in explainability research, specifically, the trade-off between real-time utility and explanation richness. Coaches working with small groups tended to prioritise immediate feedback, whereas those managing larger teams accepted post-session review. This divide parallels the computational trade-offs identified in multimodal XAI systems [3], where model-agnostic methods like SHAP or LIME can be too slow for live deployment, and attention-based mechanisms or intrinsic prototypes offer faster, if less detailed, interpretability. These qualitative insights thus provide design guidance for later implementation: real-time views may rely on lightweight, built-in attention or rule-based explanations, while post-session analysis can incorporate more computationally intensive model-agnostic methods.

The consistent selection of mobile devices (particularly tablets and smartphones) underscores the situational constraints of aquatic environments. From an interpretability standpoint, this highlights the importance of visual economy: explanations must convey meaning in compact visual forms without requiring extended analysis time. As noted in medical and sports XAI applications, interpretability is partly a function of interface bandwidth, the amount of attention and time a user can dedicate to understanding a result [2]. Designing for limited display space and rapid comprehension, therefore, becomes a practical extension of the explainability problem.

The biomechanical feedback section of the questionnaire provides a direct bridge between sensor-level interpretability and domain-level insight. Coaches rated arm-pull efficiency, body alignment, and start or turn performance as the most important indicators across all strokes. These preferences closely align with findings from attention-based HAR models such as AttnSense [8] and sensor-level attribution studies [12], which identified gyroscope and accelerometer data related to limb rotation and propulsion as key explanatory features. This overlap suggests that the model's internal logic, if designed to highlight these variables, would naturally correspond to the parameters coaches already deem most meaningful. Such alignment between human and algorithmic reasoning is central to trustworthy explainability, indicating that SWIM-360's future sensor models should emphasise interpretive coherence over novelty.

As outlined in Figure 1, the forthcoming SWIM-360 prototype will operationalise the multimodal fusion and hybrid explainability pipeline to deliver post-training visuals and interpretable insights to coaches.

Open-ended responses further reveal the epistemic function of explainability for coaches. Participants repeatedly mentioned the difficulty of observing all swimmers, the challenge of delivering timely corrections, and the need for objective recovery metrics. These statements illustrate that explainability is not merely about validating AI predictions but about extending human perception, making invisible aspects of performance (such as underwater motion or physiological recovery) visible and actionable. This interpretation resonates with the broader definition of explainable systems as cognitive extensions of human expertise rather than post-hoc justification tools [3].

The emphasis on ease of use and integrated, shareable feedback reflects the socio-technical dimension of explainability. Trust and adoption depend not only on what the model explains but also on how seamlessly the explanation fits within existing coaching practices [14]. As highlighted by both AIX360 and Rodis et al. [3], successful explainability requires careful orchestration between algorithmic, visual, and interactive components. In SWIM-360, this orchestration will involve linking attention-derived sensor weights, visual saliency overlays, and contextual natural-language summaries within a unified, coach-facing interface.

In the next development phase, SWIM-360 will integrate IMU, NIRS, and video data streams using attention-based fusion. Explainability outputs (e.g., SHAP values, attention weights, and visual saliency maps) will be linked to a coach-facing dashboard derived from the preferences reported here. A preliminary prototype will be evaluated for interpretability and usability.

6. Conclusions

This paper presented early findings from the SWIM-360 project, an initiative exploring how explainable multimodal sensing can enhance swimming performance analysis. Building on theoretical frameworks of explainable artificial intelligence, the study designed and deployed a qualitative questionnaire to capture coaches' perspectives on feedback, visualisation, and interpretability.

The results indicate that coaches place strong value on transparency and contextual understanding, preferring visual explanations such as video highlights and colour-coded overlays that clarify why a result occurs rather than simply presenting data. They also emphasise the importance of customisable and adaptive interfaces, real-time feedback where feasible, and integration of both biomechanical and physiological indicators. These preferences align closely with the multimodal explainability methods discussed in earlier sections, supporting the relevance of attention-based fusion, hierarchical prototypes, and taxonomy-guided design approaches for future system development.

These findings underline that explainability is a foundational usability feature, not a secondary add-on, for AI-driven coaching systems. The current study is limited by its small sample size and qualitative scope, yet it provides an empirical basis for designing multimodal explainable feedback. Future work will implement and validate the proposed architecture through prototype testing with coaches and athletes. Through this approach, the project aims to advance both the scientific and practical frontiers of explainable performance analytics in aquatic sports.

Author Contributions: Conceptualization, V.C. and R.Y.C.; methodology, R.Y.C. and D.S.; software, M.F. and D.P.; validation, R.Y.C., D.S., and M.M.; formal analysis, D.P.; investigation, R.Y.C. and M.F.; resources, V.C.; data curation, D.P.; writing—original draft preparation, R.Y.C.; writing—review and editing, V.C. and M.M.; visualization, M.F.; supervision, V.C.; project administration, V.C.; funding acquisition, V.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Xjenza Malta and the Malta Digital Innovation Authority (MDIA) under the Thematic Programme, with Grant Agreement Number DTP-2024-10. The Article Processing Charge (APC) was funded by the same grant.

Institutional Review Board Statement: The study was approved by the University Research Ethics Committee of the University of Malta (REDP form code: ICT-2025-00089, April 2025).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author(s).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
XAI	Explainable Artificial Intelligence
IMU	Inertial Measurement Unit
NIRS	Near-Infrared Spectroscopy

SHAP	SHapley Additive Explanations
LIME	Local Interpretable Model-agnostic Explanations
MXAI	Multimodal Explainable AI
DIME	Disentangled Multimodal Explanations
HAR	Human Activity Recognition
CNN	Convolutional Neural Network
GRU	Gated Recurrent Unit
PDP	Partial Dependence Plot
LRP	Layer-wise Relevance Propagation
HIPE	Hierarchical Prototype Explanations
SmO2	Muscle Oxygen Saturation
IBM	International Business Machines (Technology Company)
EO SwimBETTER	Commercial IMU sensor
TrainRed	Commercial NIRS sensor measuring SmO2

Appendix A. Questionnaire Format

We are developing a coach-facing dashboard for our two projects, DIVE/SWIM-360, a system that integrates wearable sensors, physiology monitoring, and video analysis to provide clear, explainable feedback on swimmer performance. Our goal is to create a tool that supports your coaching by giving you the right information at the right time, in a way that is both practical and easy to use.

The questionnaire consists of some short questions and should take no longer than 8–12 min to complete. All responses will be kept completely private, and we will not ask for any personal information at any stage.

Your input is extremely valuable. By sharing your thoughts and preferences, you will help us design a dashboard that reflects the real needs of coaches and ultimately make DIVE and SWIM-360 as useful and effective as possible.

Before continuing, please confirm that you understand the following: Your responses will be used solely for research and design of the DIVE/SWIM-360 coaching dashboard.

- No personal or identifying information will be collected.
- Participation is voluntary, and you may stop at any time.
- All responses will remain confidential and stored securely.
- By ticking the box below, you give your informed consent to take part in this questionnaire.

Q0. I have read and understood the above information and consent to participate in this questionnaire. (Required to continue)

- Yes (Continue)
- No (Submit Form)

Type: Multiple choice (single answer)

Reason: It is important to obtain consent from a coach and to ensure the coach understands what the form is used for.

Appendix A.1. General Information

Q1. What level of swimmers do you primarily coach?

- Age-group
- National
- Masters
- Other: _____

Type: Checkboxes (allow multiple selections)

Reason: Helps tailor dashboard insights and feedback complexity to the coaching level, since data needs and feedback styles differ between age groups, national, and masters swimmers.

Q2. Which country do you currently primarily coach in?

Type: Short Answer Text

Reason: Different countries may have different focuses on what to aim for in swimming.

Q3. How many swimmers do you typically monitor per session?

- 1–4
- 5–10
- 11–20
- More than 20

Type: Multiple choice (single answer)

Reason: Determines scalability needs for the dashboard (e.g., number of swimmers tracked simultaneously, level of data aggregation required).

Q4. What type of feedback from a system like our project would be most useful for your coaching? (Tick all that apply)

- Stroke technique (e.g., timing, balance, symmetry, efficiency of movement)
- Physical condition (e.g., fatigue levels, recovery status, breathing/oxygen use)
- Performance metrics (e.g., speed, stroke rate, turns, overall efficiency)
- Injury prevention signs (e.g., early indicators of stress or overuse)
- Other: _____

Type: Checkboxes (allow multiple selections)

Reason: Identifies which performance domains (technique, physiology, performance, injury prevention) should be prioritised in dashboard design and data visualisation.

Q5. Rank the following in order of importance for coaching feedback:

- Stroke efficiency
- Fatigue and recovery
- Injury prevention
- Race performance (speed, times, turns)

Type: Multiple-choice grid with ranking

Reason: Helps prioritise which data the dashboard should highlight most prominently.

Q6. How important is it that the system explains why a certain result or alert was generated?

- 1—Not Important
- 5—Very Important

Type: Linear scale (1–5 for each item in a grid)

Reason: Directly tests the importance of XAI for coaches.

Appendix A.2. Visualisation Preferences

Example of Video Replay with highlights.

In addition to showing the video, additional information is shown on top, such as the distance to go, etc.

(Example Image Showing Video Replay)

Q7. How important is it that the system explains why a result or alert appears? (1 = Not Important 5 = Very Important)

- 1—Not Important
- 5—Very Important

Type: Linear scale (1–5 for each item in a grid)

Reason: Measures how much value coaches place on interpretability and transparency in visual outputs, guiding how much explanation or justification should accompany system feedback.

Q8. If you could choose only one main type of visualisation to discuss performance, which would help you most?

- Video replay with highlights
- Graphs or trend lines
- Colour-coded visuals (e.g., balance, fatigue)
- Side-by-side comparisons (past vs. elite swimmers)

Type: Multiple choice (single answer)

Reason: Identifies preferred explanation format.

Q9. If you could pick one main screen to use every day, which would you find most useful?

- Summary dashboard (quick overview with key alerts and performance trends)
- Stroke technique view (detailed breakdown of how the stroke is performed)
- Swimmer condition view (fatigue, recovery, and early signs of injury risk)
- Customisable view (you choose what's shown, depending on the session focus)

Type: Multiple choice (single answer)

Reason: Reveals what coaches want to see first when opening the dashboard.

Q10. Rate how helpful each type of visual would be for your coaching

- Video with highlights (e.g., showing stroke mistakes or movement issues)
- Side-by-side comparisons (e.g., your swimmer vs. an elite swimmer, or vs. their past sessions)
- Graphs and trend lines (e.g., showing changes in speed, stroke rate, or lap times over time)
- Colour-coded visuals (e.g., showing imbalance, fatigue, or weak areas)

Type: Linear scale (1–5 for each item in a grid)

Reason: Gauges the usefulness of visual explanation elements for dashboard design.

Appendix A.3. Practical Use

Q11. Preferred data format:

- Simple numbers (e.g., stroke rate = 52 spm)
- System interpretation (e.g., "stroke rate below optimal range")
- Both side by side

Type: Multiple choice

Reason: Determines if coaches want raw data, interpretation, or a mix.

Q12. At first glance, how much information would you like to see?

- Key highlights only (with option for details)
- Mix of highlights + some details
- Full detailed data immediately

Type: Multiple choice

Reason: Determines information density preferences (avoiding overload).

Q13. How important is getting feedback during a training session compared to reviewing it afterwards?

- Essential (I want feedback immediately while the swimmer is training)
- Important, but post-session review is enough for me.
- Not necessary

Type: Multiple choice

Reason: Helps set system requirements for real-time vs offline dashboards.

Q14. Which devices would you most likely use the dashboard on?

- Laptop/PC
- Tablet
- Smartphone
- Poolside large screen (visible to swimmers and coaches)

Type: Checkboxes

Reason: Informs responsive design and context-of-use requirements.

Q15. How often would you realistically use such a dashboard during training or competition preparation?

- Every training session
- A few times a week during training
- Once a week during training
- Less than once a week during training
- Primarily before competitions (for race preparation and analysis)

Type: Multiple choice

Reason: Helps scope usage frequency and design priorities.

Appendix A.4. Technique and Biomechanical Feedback

In this section, we are looking to identify the key features which should be identified for swimmers when providing feedback on their technique across all strokes, turns, underwater streamline and dives.

Q16. For each of the following, how important would this type of feedback be across any stroke? (1 = Not Important 5 = Very Important)

- Body alignment and rotation
- Kick timing and symmetry
- Breathing timing/head lift
- Arm pull depth/efficiency
- Glide or underwater phase duration
- Start, turn, and push-off efficiency

Type: Linear scale (1–5 for each item in a grid)

Reason: Determines which biomechanical metrics are most relevant across strokes, ensuring the system focuses on the feedback elements that have the highest practical value for coaches.

Q17. If you could identify only three key elements to monitor in the dashboard, what would they be?

Type: Paragraph (open text)

Reason: Helps identify the highest-priority data points for inclusion, revealing which metrics coaches value most for everyday decision-making and training impact.

Q18. What is the biggest challenge you face when analysing swimmer performance without technology (e.g., video, sensors, data tools)?

Type: Paragraph (open text)

Reason: Gathers qualitative feature requests directly from coaches.

Q19. What would make DIVE/SWIM-360 most useful in your day-to-day coaching?

Type: Paragraph (open text)

Reason: Identifies the main gaps that SWIM-360 could fill.

Appendix B. Questionnaire Results

8 responses

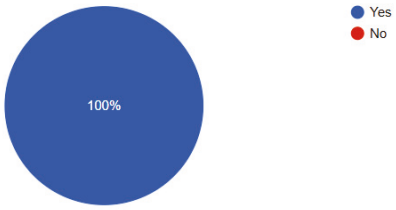
Link to Sheets

Summary Question Individual

Consent

I have read and understood the above information and consent to participate in this questionnaire.
(Required to continue)

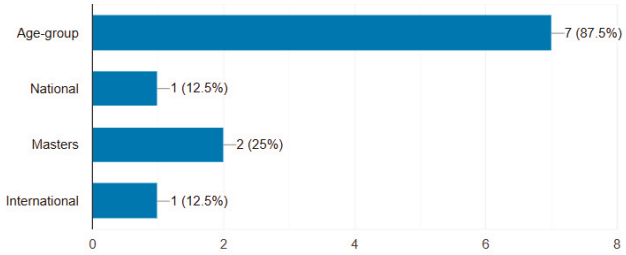
8 responses



Section 1: General Information

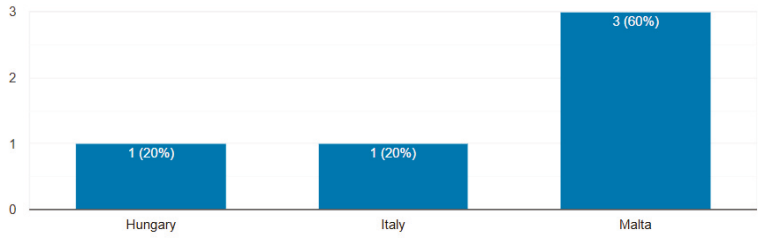
What level of swimmers do you primarily coach?

8 responses



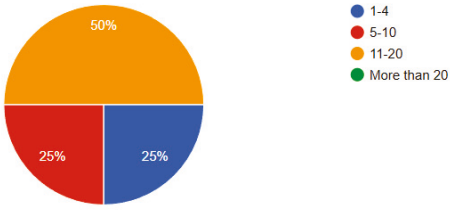
Which country do you currently primarily coach in?

5 responses



How many swimmers do you typically monitor per session?

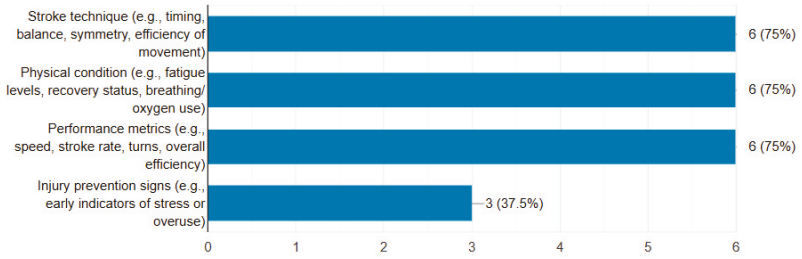
8 responses



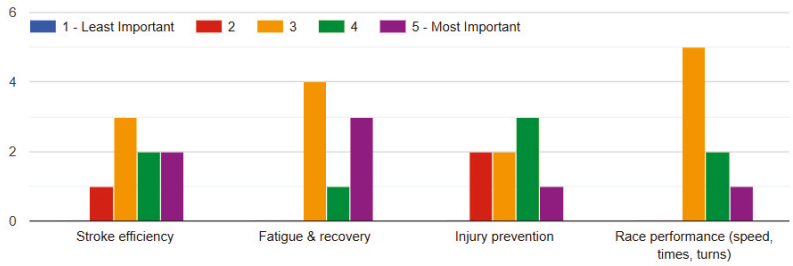
What type of feedback from a system like our project would be most useful for your coaching?

(Tick all that apply)

8 responses

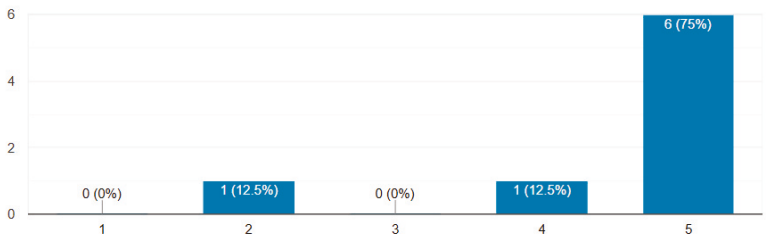


Rank the following in order of importance for coaching feedback:



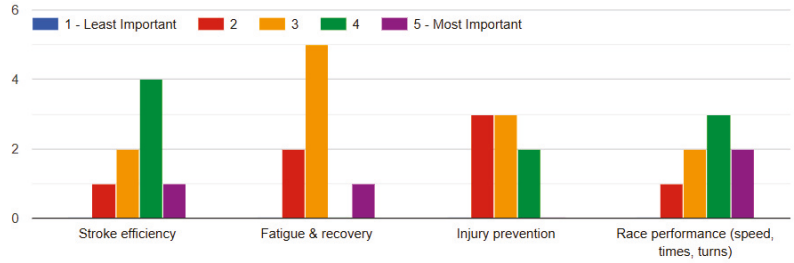
How important is it that the system explains why a certain result or alert was generated?

8 responses



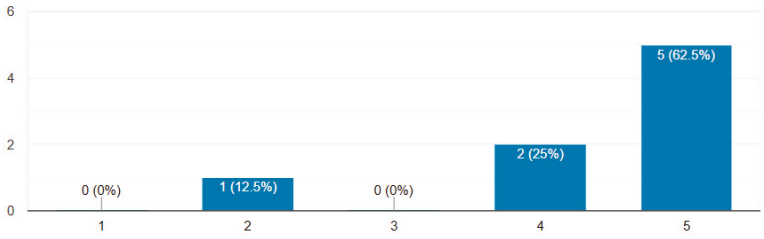
Section 2: Visualisation Preferences

Rank the following in order of importance for coaching feedback
(1 = Least Important 5 = Most Important)



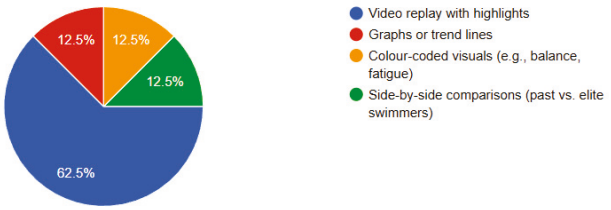
How important is it that the system explains *why* a result or alert appears?
(1 = Not Important 5 = Very Important)

8 responses



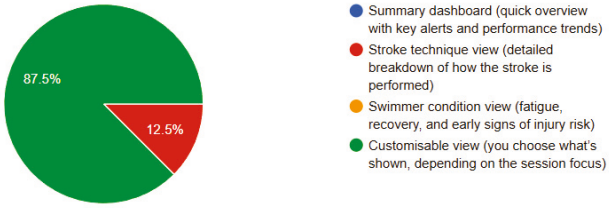
If you could choose only one main type of visualisation to discuss performance, which would help you most?

8 responses

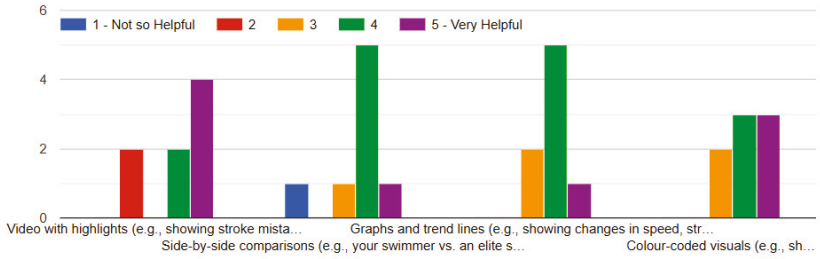


If you could pick one main screen to use every day, which would you find most useful?

8 responses



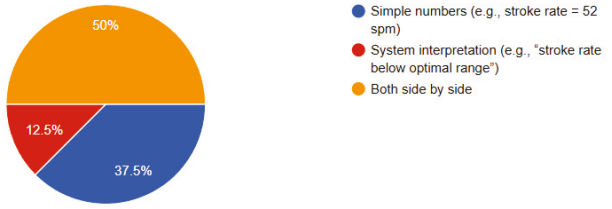
Rate how helpful each type of visual would be for your coaching



Section 3: Practical Use

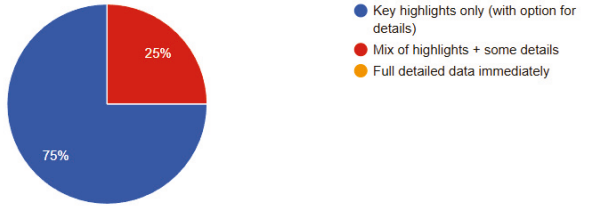
Preferred data format:

8 responses



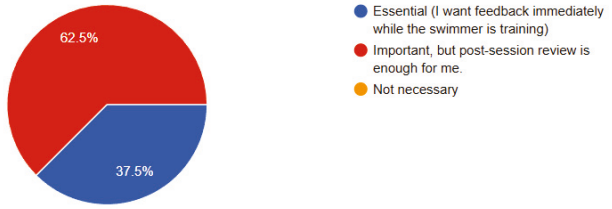
At first glance, how much information would you like to see?

8 responses



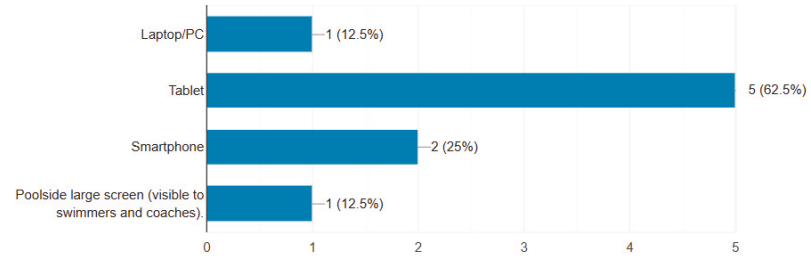
How important is getting feedback during a training session compared to reviewing it afterward?

8 responses



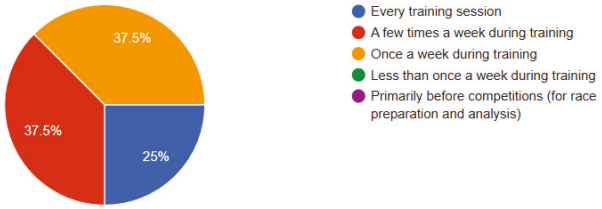
Which devices would you most likely use the dashboard on?

8 responses



How often would you realistically use such a dashboard during training or competition preparation?

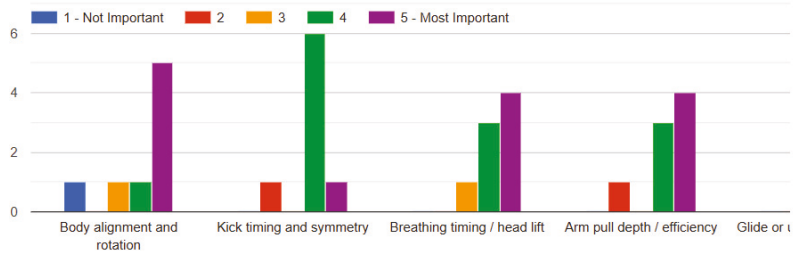
8 responses



Section 4: Technique & Biomechanical Feedback

For each of the following, how important would this type of feedback be across *any* stroke?

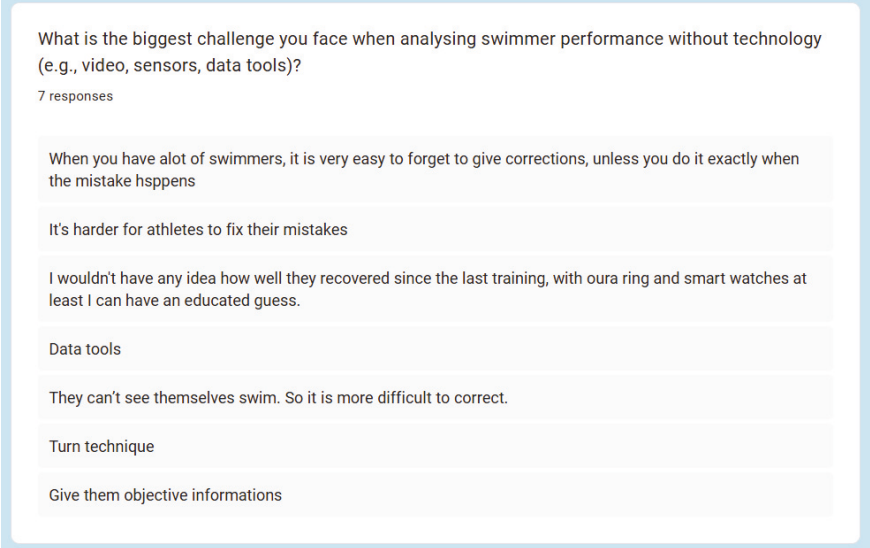
(1 = Not Important 5 = Very Important)

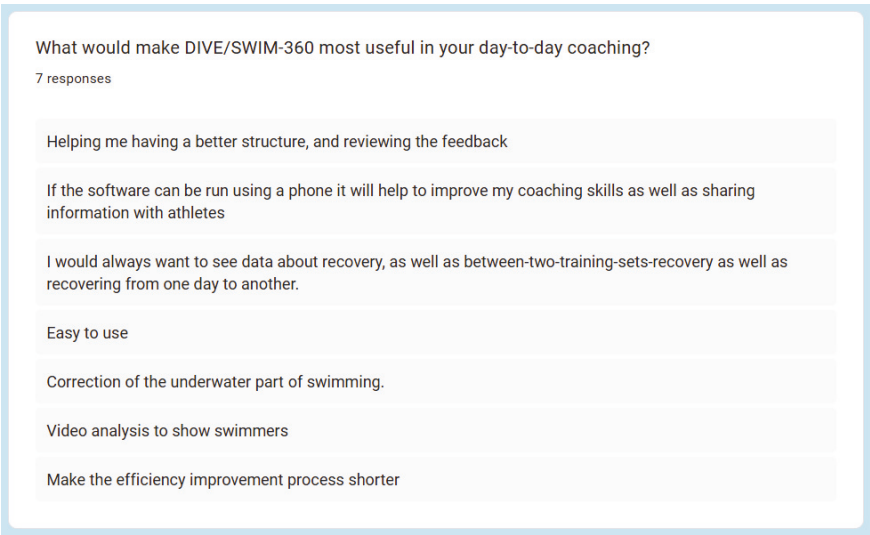


If you could identify only three key elements to monitor in the dashboard, what would they be?

7 responses

Body positioning, start and turn, breathing
Heart Rate , spm , statistics on past sessions to compare with new session
Stroke timing, frequency, underwater efficiency
Body alignment, UW, arm pull efficiency
Body position, stroke efficiency, timing
Stroke efficiency, body alignment, body position
Stroke rate, distance per stroke, energy cost (heart rate or lactate)





References

1. Gulshad, S.; Long, T.; van Noord, N. Hierarchical Explanations for Video Action Recognition. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Vancouver, BC, Canada, 17–24 June 2023; pp. 3703–3708. [CrossRef]
2. Tjoa, E.; Guan, C. A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. *IEEE Trans. Neural Netw. Learn Syst.* **2021**, *32*, 4793–4813. [CrossRef] [PubMed]
3. Rodis, N.; Sardianos, C.; Radoglou-Grammatikis, P.; Sarigiannidis, P.; Varlamis, I.; Papadopoulos, G.T. Multimodal Explainable Artificial Intelligence: A Comprehensive Review of Methodological Advances and Future Research Directions. *IEEE Access* **2024**, *12*, 159794–159820. [CrossRef]
4. Arya, V.; Bellamy, R.K.E.; Chen, P.-Y.; Dhurandhar, A.; Hind, M.; Hoffman, S.C.; Houde, S.; Liao, Q.V.; Luss, R.; Mojsilović, A.; et al. AI Explainability 360: An Extensible Toolkit for Understanding Data and Machine Learning Models. *J. Mach. Learn. Res.* **2020**, *21*, 130:5114–130:5119.
5. Saha, A.; Gupta, S.; Ankireddy, S.K.; Chahine, K.; Ghosh, J. Exploring Explainability in Video Action Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–18 June 2024.
6. Zhi, Y.; Tong, Z.; Wang, L.; Wu, G. MGSampler: An Explainable Sampling Strategy for Video Action Recognition. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 1493–1502. [CrossRef]
7. Mahmud, S.; Tanjid Hasan Tonmoy, M.; Kumar Bhaumik, K.; Mahbubur Rahman, A.K.M.; Ashraful Amin, M.; Shoyaib, M.; Asif Hossain Khan, M.; Ahsan Ali, A. Human Activity Recognition from Wearable Sensor Data Using Self-Attention. In *Frontiers in Artificial Intelligence and Applications*; IOS Press: Amsterdam, The Netherlands, 2020. [CrossRef]

8. Ma, H.; Li, W.; Zhang, X.; Gao, S.; Lu, S. AttnSense: Multi-Level Attention Mechanism For Multimodal Human Activity Recognition. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, Macao, China, 10–16 August 2019; International Joint Conferences on Artificial Intelligence Organization: Macao, China, 2019; pp. 3109–3115. [CrossRef]
9. Lattanzi, E.; Calisti, L.; Capellacci, P. Lightweight Accurate Trigger to Reduce Power Consumption in Sensor-Based Continuous Human Activity Recognition. *Pervasive Mob. Comput.* **2023**, *96*, 101848. [CrossRef]
10. Vijay Arya. Trusted-AI/AIX360. 2025. Available online: https://github.com/Trusted-AI/AIX360 (accessed on 28 August 2025).
11. Sandamal, K.; Arachchi, S.; Erkudov, V.O.; Rozumbetov, K.U.; Rathnayake, U. Explainable Artificial Intelligence for Fitness Prediction of Young Athletes Living in Unfavorable Environmental Conditions. *Results Eng.* **2024**, *23*, 102592. [CrossRef]
12. Navakauskas, D.; Dumpis, M. Wearable Sensor-Based Human Activity Recognition: Performance and Interpretability of Dynamic Neural Networks. *Sensors* **2025**, *25*, 4420. [CrossRef] [PubMed]
13. LinLin, H.; Sangheang, L.; GuanTing, S. CAM-Vtrans: Real-Time Sports Training Utilizing Multi-Modal Robot Data. *Front. Neurorobot.* **2024**, *18*, 1453571. [CrossRef] [PubMed]
14. Almasi, S.; Bahaadinbeigy, K.; Ahmadi, H.; Sohrabei, S.; Rabiei, R. Usability Evaluation of Dashboards: A Systematic Literature Review of Tools. *BioMed Res. Int.* **2023**, *2023*, 9990933. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

sensors

MDPI

Article

Generative Sign-Description Prompts with Multi-Positive Contrastive Learning for Sign Language Recognition

Siyu Liang [1,2,3], Yunan Li [1,2,3], Wentian Xin [4], Huizhou Chen [1,2,3], Xujie Liu [1,2,3], Kang Liu [1,2,3] and Qiguang Miao [1,2,3,*]

[1] School of Computer Science and Technology, Xidian University, Xi'an 710071, China; syliang_233@stu.xidian.edu.cn (S.L.); yunanli@xidian.edu.cn (Y.L.); huizhouchen@stu.xidian.edu.cn (H.C.); 24031110049@stu.xidian.edu.cn (X.L.); kangliu@stu.xidian.edu.cn (K.L.)
[2] Xi'an Key Laboratory of Big Data and Intelligent Vision, Xi'an 710071, China
[3] Key Laboratory of Collaborative Intelligence Systems, Ministry of Education, Xidian University, Xi'an 710071, China
[4] College of Information Science and Technology, Dalian Martime University, Dalian 116026, China; wtxin@dlmu.edu.cn
* Correspondence: qgmiao@xidian.edu.cn

Abstract

While sign language combines sequential hand motions with concurrent non-manual cues (e.g., mouth shapes and head tilts), current recognition systems lack multimodal annotation methods capable of capturing their hierarchical semantics. To bridge this gap, we propose GSP-MC, the first method integrating generative large language models into sign language recognition. It leverages retrieval-augmented generation with domain-specific large language models and expert-validated corpora to produce precise multipart descriptions. A dual-encoder architecture bidirectionally aligns hierarchical skeleton features with multi-level text descriptions (global, synonym, part) through probabilistic matching. The approach combines global and part-level losses with KL divergence optimization, ensuring robust alignment across relevant text-skeleton pairs while capturing sign semantics and detailed dynamics. Experiments demonstrate state-of-the-art performance, achieving 97.1% accuracy on the Chinese SLR500 (surpassing SSRL's 96.9%) and 97.07% on the Turkish AUTSL (exceeding SML's 96.85%), confirming cross-lingual potential for inclusive communication technologies.

Keywords: sign language recognition; contrastive learning; generative large language model; modality fusion

1. Introduction

Due to its importance in bridging communication gaps across diverse human communities, sign language recognition (SLR) has attracted substantial research interest. The advent of high-precision depth sensors, including Kinect [1] and RealSense [2], coupled with advances in pose estimation algorithms [3,4], has significantly simplified the acquisition of human joint positions in recent years. Skeleton-based SLR methods, which rely on body joint movements, have gained considerable attention due to their computational efficiency and robustness to variations in lighting conditions, viewpoints, and background noise. An increasing number of SLR works are based on the skeleton modality [5] or use skeletons as one of the multimodal inputs [6–9].

The remarkably rapid development of pre-trained generative LLMs has led to their expanding applications across various recognition domains [10–12], particularly in action

recognition. Approaches such as GAP [13] and SpaMo [14] have demonstrated how LLMs can enhance recognition by generating fine-grained textual descriptions or interpreting spatial-motion features. Although these methods primarily employ LLMs as sophisticated text processors, their generative capabilities remain underexplored for domain-specific applications like sign language recognition. This presents several unique challenges, as sign language processing requires both linguistic precision and specialized domain knowledge that general-purpose LLMs typically lack. Therefore, potential solutions must address the fundamental tension between domain-specific motion accuracy and linguistic expressiveness in generated descriptions.

Although prompt engineering has enabled LLMs to assist in action recognition by generating auxiliary text descriptions, such approaches face unique hurdles in SLR. Sign language requires expert knowledge, as subtle variations in hand shape, motion, or expression convey different meanings. General LLMs often produce descriptions that are either overly generic or semantically inconsistent. Existing methods [13,15], designed for action recognition, struggle with inaccuracies and hallucinations in sign descriptions. This limitation calls for domain-aware prompting techniques that harmonize LLMs' generative flexibility with the structural precision demanded by sign language. This suggests the need for new prompting paradigms that can bridge the generative capacity of LLMs with the strict descriptions of expert-defined signs.

Contrastive learning has revolutionized unsupervised representation learning in multiple domains. MoC [16] implements a momentum-based dynamic dictionary with a queue and moving-average encoder for scalable contrastive learning. SimCL [17] proposes a method for text-to-image models, treating multiple images generated from the same text prompt as positive samples. MC [18] proposes a multi-view enhanced contrastive learning method for visual representation by maximizing agreements between multi-view radiographs and their corresponding reports, solving medical imaging issues. Sign language recognition challenges contrastive learning with many-to-many relationships: One sign has multiple valid descriptions (all of which should be treated as positives), and descriptions often focus on partial actions. The single positive contrastive learning method struggles with these variable, incomplete positives, requiring new approaches that handle probabilistic alignments while preserving discrimination.

The contributions of our proposed method are as follows.

1. To the best of our knowledge, we are the first to integrate generative LLMs into SLR through our Generative Sign-description Prompts (GSP). GSP employs retrieval-augmented generation with domain-specific LLMs to produce accurate multipart sign descriptions, providing reliable textual grounding for contrastive learning.
2. We design the Multi-positive Contrastive learning (MC) approach, which combines retrieval-augmented generative descriptions from expert-validated Knowledge Bases and a novel multi-positive contrastive learning paradigm.
3. Comprehensive experiments on the Chinese SLR500 and Turkish AUTSL datasets further validate the effectiveness of our method, achieving state-of-the-art accuracy (97.1% and 97.07%, respectively). The consistent and robust performance across languages demonstrates generalization capabilities.

2. Related Work

2.1. Skeleton-Based Sign Language Recognition

Current research in skeleton-based SLR has explored several promising approaches. Self-supervised learning approaches, such as Sign-BERT [19] and BEST [7], employ masked sequence modeling and quantized pseudo-labels to learn sign representations from hand pose sequences. These methods demonstrate the potential for learning sign representations

without extensive annotations, although they primarily focus on low-level motion patterns rather than linguistic semantics. Graph-based architectures like CoSign [6] utilize specialized graph convolutional networks to capture the skeletal dynamics but often neglect the global semantic relationships inherent in sign language. Meanwhile, multimodal pretraining methods such as MASA [8] integrate motion-aware autoencoders with semantic alignment to learn joint visual–linguistic representations. Despite these advances, existing methods exhibit persistent limitations by treating sign language primarily as motion sequences while neglecting its inherently structured linguistic nature. This inevitably leads to excessive dependence on data-driven motion patterns, weak integration with linguistic knowledge, and inadequate visual–semantic alignment, which are fundamental shortcomings that hinder linguistically grounded SLR systems. Thus, our work specifically addresses these gaps by multi-positive contrastive learning, aligning each sign with multiple expert-verified descriptions while maintaining cross-modal discrimination via fixed-text-encoder training. This approach learns robust visual features that absorb natural sign variations without computational overhead during inference, achieving more linguistically grounded recognition than previous methods.

2.2. Text-Augmented Sign Language Representation

Recent progress in combining textual and visual features for SLR predominantly adopts two distinct strategies. On the one hand, manual text encoding methods such as NLA-SLR [20] and C^2ST [21] leverage predefined gloss vocabularies but are constrained by static description sets that fail to adequately capture execution variations and face scalability issues due to high manual annotation costs. On the other hand, generative text integration approaches, including Action-GPT [22] and GAP [13], employ LLM-based action description generation. However, these primarily target generic action recognition or vision-language tasks, leaving key SLR challenges unaddressed: hallucinated sign descriptions, lack of sign-specific linguistic constraints, and inherent difficulties in modeling simultaneous sign components (e.g., hand shape, motion, and orientation). While Sign2GPT [23] and Action-GPT [22] pioneered LLMs for sign/action description, GSP-MC introduces two novel contributions: (1) generative LLMs with RAG-enhanced domain knowledge to produce hallucination-free, discriminative sign descriptions (unlike prior LLMs used as text processors), and (2) multi-positive contrastive learning explicitly modeling the many-to-many relationships between partial skeleton motions and their valid descriptions. This approach uniquely addresses sign language's structural semantics where existing single-positive contrastive frameworks fail.

3. Materials and Methods

The Generative Sign-description Prompts with the multi-positive contrastive learning (GSP-MC) method augments skeleton-based sign language recognition through multimodal representation learning. As illustrated in Figure 1, the architecture comprises three elements: (1) a group-specific skeleton encoder extracting hierarchical motion features, (2) a Generative Sign-description Prompt (GSP) module producing expert-knowledge-grounded text descriptions, and (3) a multi-positive contrastive learning mechanism (MC) aligning visual and textual representations. The method maintains computational efficiency during inference using only the skeleton encoder.

The skeleton encoder E_s extracts both the global skeleton features $\mathbf{S}_g$ and the part-specific features $\mathbf{S}_p$ from the input pose sequences. The GSP employs LLMs to produce descriptive texts t, which are subsequently encoded into text features $\mathbf{T}$ by the text encoder E_t. These multimodal features are then optimized through our proposed multi-positive

contrastive learning approach, which enhances the model's capacity to capture fine-grained action semantics from textual descriptions.

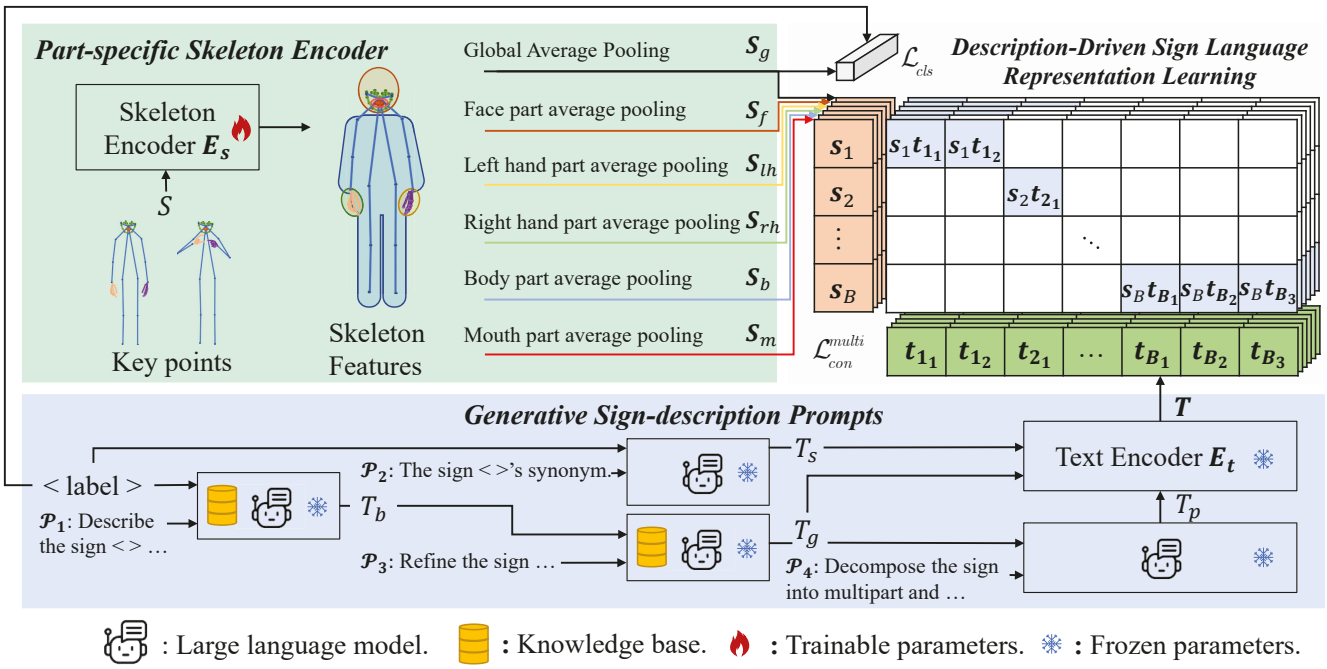


Figure 1. The overall architecture of the GSP-MC method. $\mathbf{S}_g$ denotes global skeleton features for classification; $\mathbf{S}_f$, etc., represents part-specific features for contrastive learning, and $\mathbf{T}$ is encoded text features. Training is guided by $\mathcal{L}_{cls}$ and $\mathcal{L}_{con}$ losses.

3.1. Part-Specific Skeleton Encoder

Our model processes input skeleton sequences $S \in \mathbb{R}^{B\times 3\times N\times T}$ as input, where B is the batch size, 3 represents (x, y, confidence), N denotes the number of joints, and T indicates the temporal sequence length. The model predicts the labels $l \in \mathbb{R}^{B1}$ as output.

Keypoint Selection and Grouping. Using HR-Net [4] as our skeleton estimator, we extract 87 keypoints per frame, strategically divided into five anatomical groups: 15 for the body S_b, 21 for each hand S_{lh} and S_{rh}, 10 for the mouth S_m , and 20 for the face S_f. We provide a detailed analysis of this selection strategy in Section 4.3.3.

Skeleton Encoder. Representing the skeleton as a graph $S = \{V, E\}$ with joints V and edges E, we process each part P through layered graph convolutions.

$$\mathbf{S}_{P,l+1} = \sigma\left(\mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}\mathbf{S}_{P,l}\Theta_l\right) \tag{1}$$

where $\mathbf{D} \in \mathbb{R}^{N\times N}$ is the degree matrix. $\mathbf{A}$ is the adjacency matrix. $\Theta_l \in \mathbb{R}^{C_l\times C_{l+1}}$ is the learnable parameter at layer l. And σ is the activation function.

Our basic block employs multiple CTR-GC blocks [24], with each block integrating a Channel-wise Topology Refinement Graph Convolution layer. We fuse channel-wise topology refinement with graph convolution to yield the final part-specific representation.

Skeleton Classification. The model optimizes a standard cross-entropy loss:

$$\mathcal{L}_{cls} = -\mathbf{Y}\log p_\theta(\mathbf{S}_g) \tag{2}$$

where $\mathbf{Y}$ denotes ground truth labels. $\mathbf{S}_g$ represents the global skeleton features. And $p_\theta(x)$ is the predicted probability distribution.

3.2. Generative Sign-Description Prompts

The Generative Sign-description Prompts (GSPs) establish a systematic approach for generating accurate and comprehensive sign language descriptions through advanced language model techniques. As illustrated in Figure 2, the system integrates domain-specific knowledge from authoritative sources, including official sign definitions and expert textbooks. The LLM model's DAG functionality, which we employ for retrieval-augmented generation, supports Knowledge Bases up to 1GB and individual documents up to 50 MB, providing ample capacity for our requirements.

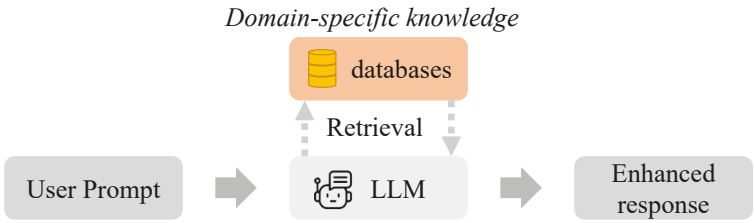


Figure 2. Architecture of the customized LLM for sign description generation.

3.2.1. Expert Dataset Construction and Standardization

The method addresses the inherent challenges of professional sign language descriptions through a dual-path generation mechanism. Professional sign language definitions frequently employ substituted descriptions that create barriers for machine interpretation, manifesting as standardized manual alphabet references ("uses the '5' handshape") or cross-sign equivalencies ("identical to the sign 'good'"). To overcome this, the system implements a rigorous standardization process where primary descriptions T_b are generated through RAG-augmented generation anchored in authoritative sources:

$$T_b = \text{LLM}_{\text{RAG}}(\text{label}, \mathcal{P}_1) \tag{3}$$

Simultaneously, the method incorporates controlled diversity through synonym variations T_s generated by leveraging the LLM's creative potential within designed constraints:

$$T_s = \text{LLM}(\text{label}, \mathcal{P}_2) \tag{4}$$

The prompt templates $\mathcal{P}_{1,2}$ incorporate domain-specific instruction tuning to ensure outputs maintain both high factual accuracy through expert grounding and rich linguistic diversity. For instance, a base description "palm pushes forward" might be expanded to "arm extends with palm facing outward" while preserving essential core features. This dual method effectively balances the need for standardization against the requirement for varied expressions to enhance model robustness.

3.2.2. Redundancy Elimination and Multi-Part Decomposition

The method implements advanced processing to address extraneous information that is common in professional sign language materials. Through targeted prompt design $\mathcal{P}_4$, the system automatically filters irrelevant content such as homophonic explanations ("'One' is homophonous with 'idea'"), focusing exclusively on action-relevant descriptions. As shown in Table 1, the refinement process transforms initial outputs T_b into complete alignment results comparing formulations T_g while preserving expert-validated knowledge.

$$T_g = \text{LLM}_{\text{RAG}}(T_b, \mathcal{P}_3) \tag{5}$$

Furthermore, the system introduces innovative multi-part decomposition, automatically generating part-specific texts T_p with corresponding anatomical annotations.

Table 1. Alignment results comparing primary T_b and refined T_g descriptions, highlighting: sign label, substitute descriptions, expert-validated knowledge, primary descriptions T_b, and refined descriptions T_g.

Primary Descriptions T_b	Expert-Validated Knowledge	Refined Descriptions T_g
Devoted: (I) Make the sign for "love". (II) Extend the thumb with one hand and place it on the palm of the other hand, then raise it upwards.	The sign for "love": Gently caress the back of the thumb with one hand, expressing a feeling of "tenderness".	Devoted: (1) Gently caress the back of the thumb with one hand, expressing a feeling of "tenderness". (2) Extend your thumb with one hand, sit on the other palm, and lift it up.
Ambience: (1) One hand forms the manual sign "Q", with the fingertips pointing inward, placed at the nostrils. (2) Extend your index finger with one hand and make a big circle with your fingertips facing down.	The manual sign "Q": One hand with the right thumb down, the index and middle fingers together on top, the thumb, index, and middle fingers pinched together, the fingertips pointing forward and slightly to the left, the ring and little fingers bent, the fingertips touching the palm.	Ambience: (1) One hand with the right thumb down, the index and middle fingers together on top, the thumb, index, and middle fingers pinched together, the fingertips pointing forward and slightly to the left, the ring and little fingers bent, the fingertips touching the palm. The fingertips are pointing inward, placed at the nostrils. (2) Extend index finger with one hand and make a circle with your fingertips facing down.

$$T_p = \mathrm{LLM}(T_g, \mathcal{P}_4) \tag{6}$$

The structured output T_p pairs each text segment (T_{p1}, T_{p2}, etc.) with its corresponding sign language components. As demonstrated in Figure 3, T_{p1} describes the movement involving both $part_1$ and $part_2$, while T_{p2} relates to $part_2$ and $part_3$, which establishes a many-to-many mapping between text and human body parts.

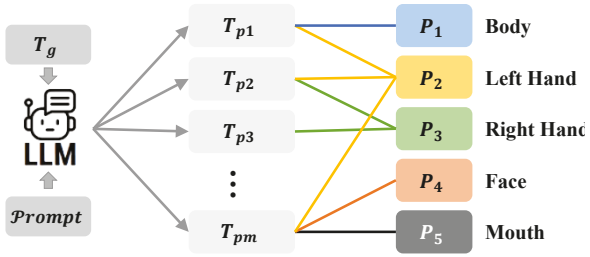


Figure 3. Generation of multipart texts.

3.2.3. Text Encoder

The text encoding component employs the pre-trained CLIP text encoder as its backbone E_t, processing three distinct text modalities: global descriptions T_g, semantically rich synonym variants T_s, and fine-grained part-specific texts T_p. The encoding process begins with standard tokenization, followed by subsequent processing through 12 transformer blocks with multi-head self-attention mechanisms. This architecture ultimately generates context-aware hierarchical linguistic representations that are efficiently aggregated into fixed-dimensional feature vectors $\mathbf{T_g}$, $\mathbf{T_s}$, and $\mathbf{T_p}$.

For part-specific encoding, individual body part descriptions undergo independent processing through identical architectural components, ensuring feature space consistency across all modalities. The use of frozen pre-trained weights maintains established linguistic

knowledge while providing computational efficiency during training. This design choice allows the method to focus innovation on the description generation aspects while leveraging proven text encoding capabilities.

3.3. Description-Driven Sign Language Representation Learning

3.3.1. Text-Conditioned Multi-Positive Alignment

Our method uses a dual-encoder architecture. As illustrated in Figure 1, it includes a skeleton encoder E_s and a text encoder E_t. E_s processes skeletal motion data $\mathbf{S} \in \mathbb{R}^{(B\times 3\times N\times T)}$, while E_t handles action descriptions. Unlike traditional one-hot label supervision, the proposed approach uses natural language descriptions, which provide richer supervision for skeleton classification. Our description-driven representation learning is detailed in Figure 4. The process consists of three key stages. First, as shown in Figure 4(1), our Generative Sign-description Prompt (GSP) module leverages Large Language Models (LLMs) to produce two types of descriptions for a sign: a holistic global description and several fine-grained part-specific descriptions focusing on individual body parts (e.g., hands, face, mouth).

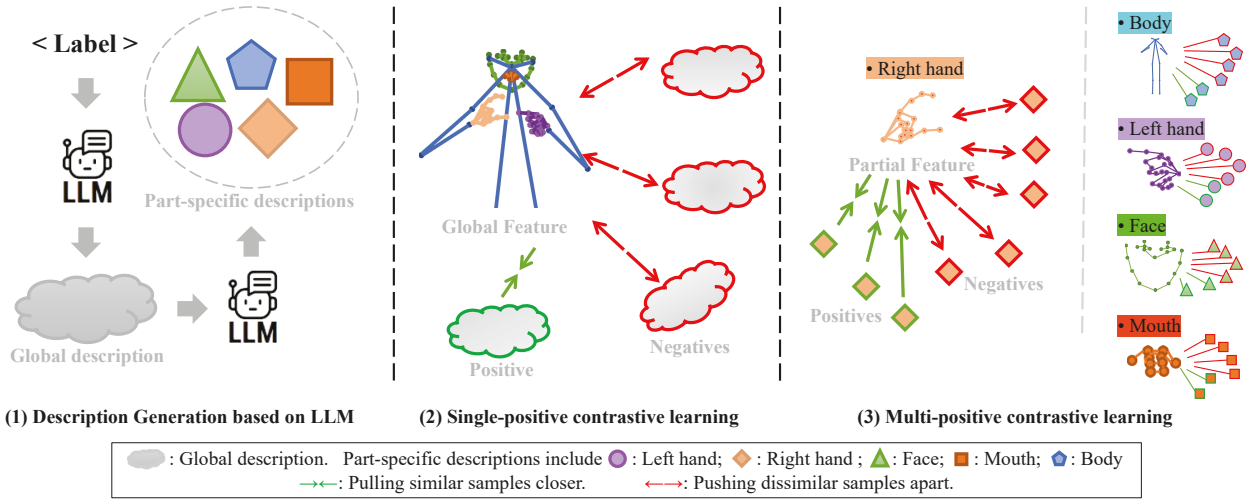


Figure 4. The mechanism of our Multi-positive Contrastive Learning. The process progresses from (**1**) LLM-based generation of global and part-specific descriptions, to (**2**) conventional single-positive alignment, and culminates in (**3**) our multi-positive alignment, where one skeleton feature aligns with multiple corresponding text features.

This rich textual data enables our contrastive learning. Figure 4(2) illustrates the conventional single-positive contrastive learning scenario. Here, each skeleton sequence corresponds to one global description T_g or synonym T_s, resulting in the number of text features M matching the batch size B (i.e., $M = B$). The goal is to "pull" the skeleton's feature closer to its single positive text feature (visualized by a green arrow) while "pushing" it away from unrelated negative features (red arrows).

Our method's primary innovation is showcased in Figure 4(3), which depicts multi-positive contrastive learning. In the part-specific scenario, multiple descriptions T_p are generated for each skeleton, leading to $M > B$. Here, a single skeleton feature is aligned with multiple positive text descriptions simultaneously. As illustrated, green arrows show the feature being pulled toward all its relevant part descriptions (distinguished by shapes like diamonds for different body parts), while red arrows show repulsion from all negatives. This hierarchical representation allows the model to learn both holistic semantics and fine-grained dynamics, which is the key driver of our method's advantage.

The dual-encoders are jointly optimized by contrasting skeleton–text pairs in two directions within the batch:

$$\mathbf{q}_{ia}^{s\to t}(\mathbf{s}) = \frac{\exp(\text{sim}(\mathbf{s}_i, \mathbf{t}_{ij})/\tau)}{\sum_{i=1}^{B}\sum_{k=1}^{m_i}\exp(\text{sim}(\mathbf{s}_i, \mathbf{t}_{ik})/\tau)} \tag{7}$$

$$\mathbf{q}^{t\to s}(\mathbf{t}) = \mathbf{q}^{s\to t}(\mathbf{t})^T \tag{8}$$

where s, t are encoded features of the skeleton and the text. $a = \sum_{b=1}^{i-1} m_b + j$. j represents the index of the j-th text feature within the set of m text features that correspond to the s_i keypoint action feature. $sim(s,t)$ is the cosine similarity, τ is the temperature parameter, and B is the batch size. In contrast to image–text pairs in CLIP [25], which are one-to-one mappings, in our proposed setting, it is possible to have more than one positive matching and actions of different categories forming negative pairs.

Given multiple candidate text features for each sign language sample, we establish a probabilistic match where at least one text feature t corresponds to each skeleton feature s. The true correspondence distribution $\mathbf{p} \in \mathbb{R}^{(B,M)}$ is defined as

$$\mathbf{p}_i = \frac{\mathbb{I}_{\text{match}(s,t_i)}}{\sum_{c=1}^{M}\mathbb{I}_{\text{match}(s,t_c)}} \tag{9}$$

where the indicator function $\mathbb{I}_{\text{match}(s,t_i)}$ equals 1 when s and t_i share the same sign label and 0 otherwise. This formulation explicitly captures the multi-positive relationships between skeleton features and their corresponding text descriptions.

The contrastive learning objective employs KL divergence to align the predicted and true distributions:

$$\mathcal{L}_{\text{con}}(s,t) = \frac{1}{2}\mathbb{E}_{\mathbf{s},\mathbf{t}\sim\mathcal{D}}\Big[\text{KL}(q^{s\to t}(\mathbf{s}), \mathbf{p}) + \text{KL}(q^{t\to s}(\mathbf{t}), \mathbf{p}^T)\Big] \tag{10}$$

where D is the entire dataset.

This symmetric loss function generalizes standard single-positive contrastive learning [26], where $\mathbf{p}$ reduces to a one-hot vector. Although related to [27], our key distinction is the explicit treatment of all texts of the same label as matched positives. Optimization brings together partial descriptions of identical signs while separating them from different signs, enhancing feature discriminability.

Our multi-positive contrastive learning fundamentally differs from standard multi-view contrastive setups. While traditional approaches like CLIP enforce a strict one-to-one correspondence between modalities (e.g., one image to one caption), sign language inherently exhibits many-to-many relationships due to execution variations and multiple valid descriptions. For instance, a single sign can be correctly described from different perspectives (holistic action, specific handshape, movement trajectory). Unlike standard contrastive learning, where each sample has exactly one positive pair, our approach models multiple valid positive pairs per sign sample through the probabilistic matching distribution in Equation (9). This explicitly captures the reality that one sign can have multiple correct textual descriptions, allowing the model to learn a more robust and comprehensive representation by pulling a single skeleton feature towards multiple text-based positive anchors in the embedding space.

3.3.2. Hierarchical Part Contrastive Learning

Considering the priors of human body parts, the skeleton can be divided into multiple groups. We illustrate the overall architecture in Figure 1. We apply contrastive loss on different part features as well as global features and propose a multipart contrastive loss.

The part feature could be obtained with part pooling, where joint features within the same group are aggregated to generate a part representation. More specifically, we choose the features for part feature pooling. Our part-partition strategy is five-part partition. The body is divided into five groups: left hand, right hand, face, mouth, and body. We then engage in contrastive learning by comparing the sign-relevant partial text features obtained from Section 3.2 with their corresponding body parts.

Our dual-encoder architecture employs an asymmetric training–inference design for computational efficiency. During training, the skeleton encoder learns to align its representations with multiple text descriptions through contrastive learning, effectively distilling linguistic knowledge into visual features. The text encoder acts as a teacher, guiding the skeleton encoder to develop richer semantic representations that capture both global sign meanings and part-specific dynamics. At inference time, only the skeleton encoder is deployed, as it has already internalized the semantic knowledge from the text descriptions. This design principle ensures that our method maintains the same computational efficiency as conventional skeleton-only approaches during deployment while benefiting from the superior feature representations learned through multimodal training.

The loss function of multipart contrastive loss can be represented as follows:

$$\mathcal{L}_{con}^{multi} = \frac{1}{N}\left(\mathcal{L}_{con}(\mathbf{S}_g, \mathbf{T}_g) + \mathcal{L}_{con}(\mathbf{S}_g, \mathbf{T}_s) + \sum_{i=1}^{P} \mathcal{L}_{con}(\mathbf{S}_{pi}, \mathbf{T}_{pi})\right) \tag{11}$$

where N is the number of terms used to calculate the contrastive loss. The hierarchical approach offers distinct advantages by inherently respecting the compositional nature of sign language while enabling robust learning from partial observations. Through explicit modeling of part–whole relationships, it achieves superior generalization. As demonstrated in our experiments, this multi-granularity representation significantly outperforms global-only contrastive approaches while maintaining computational efficiency.

3.4. Composite Training Objective

Having introduced the individual components, we now formalize the complete optimization objective that jointly trains the skeleton encoder and text alignment modules. The composite training objective combines classification and contrastive losses:

$$\mathcal{L}_{total} = \mathcal{L}_{cls}(\mathbf{S}_g) + \alpha \mathcal{L}_{con}^{multi}(\mathbf{S}, \mathbf{T}) \tag{12}$$

where $\mathcal{L}_{cls}$ denotes the standard cross-entropy classification loss, $\mathcal{L}_{con}^{multi}$ represents our multi-positive contrastive loss, and α is a fixed weighting parameter balancing the two objectives. The weighting parameter α controls the relative importance of semantic alignment versus classification accuracy, which we empirically set to 0.5 based on validation.

During training, the skeleton encoder processes input poses to generate $\mathbf{S}_g$ through the global average pooling of all joint nodes, while $\mathbf{S}_p$ is obtained by average-pooling predefined groups of related joints. Both features are projected to match the text feature dimension via separate fully connected layers. The text descriptions generated by the GSP are encoded in fixed-dimensional representations $\mathbf{T}$ using the text encoder. At inference time, the method utilizes only the global features of the skeleton encoder $\mathbf{S}_g$ for the final prediction, ensuring no additional computational overhead compared to conventional skeleton-based approaches. During inference, the text encoder and part-specific branches are discarded, reducing the method to a single-stream skeleton encoder matching the efficiency of conventional approaches while retaining the benefits of multimodal training.

4. Experiments

4.1. Experimental Setup

4.1.1. Datasets

We conduct comprehensive evaluations on two large-scale datasets.

SLR-500 [28] is a Chinese sign language dataset containing 500 vocabulary items performed by 50 native signers under controlled laboratory conditions. The dataset is divided into 87,500 training videos (35 signers) and 37,500 test videos (15 signers) for signer-independent evaluation. It provides comprehensive coverage of basic Chinese signs with natural variations in signing styles.

AUTSL [29] features 226 Turkish sign language gestures collected from 43 signers, with 28,142 training and 3742 test samples. The dataset presents unique challenges due to significant inter-signer variations and diverse recording conditions. Its medium-scale vocabulary and realistic signing variations make it valuable for robustness evaluation.

4.1.2. Implementation Details

In our method, we employ HR-Net [4] to extract 87 semantically significant keypoints per frame, which our ablation studies identified as optimal for the representation of sign language. These keypoints are processed by a CTR-GCN [24] backbone with multiscale temporal convolution, preserving both spatial and temporal dynamics. The CLIP text encoder [25] processes multiple description variants, including global descriptions T_g, synonym variations T_s, and part-specific descriptions T_p, using a contrastive loss temperature parameter $\tau = 0.1$ throughout all experiments. Training protocols vary by dataset: for SLR-500 we employ 110 training epochs with batch size 120, implementing a 5-epoch linear warm-up phase followed by an initial learning rate 0.06 (gradually annealed by $\times 0.1$ at epochs 90 and 100) and weight decay 5×10^{-4}. The AUTSL dataset follows similar 110-epoch training, but with batch size 80, initial learning rate 0.04, and weight decay 4×10^{-4} while maintaining identical reduction scheduling.

Our method consists of generated content in authoritative sign language resources, including three official Chinese dictionaries [30–32] and the Turkish National Sign Language Dictionary [33], comprising approximately 8000 vocabulary entries (1 MB of text). For text generation, we utilize Moonshot https://platform.moonshot.cn/(accessed on 26 March 2025) AI models (v1-8K/v1-32K) to produce synonym variants, expert-verified descriptions, and part-specific texts. All implementations use PyTorch 1.8.1 running on A100-PCIE-40GB GPU hardware (manufactured by NVIDIA Corporation, Santa Clara, CA, USA).

4.2. Comparison with State-of-the-Art Methods

In this section, we compare our method with the existing state-of-the-art (SOTA) methods using the SLR-500 dataset. Additionally, we also conduct performance comparisons using the large-scale AUTSL sign language dataset.

4.2.1. Performance Comparison on SLR-500 Dataset

Current SOTA methods exhibit several limitations that our approach addresses. While SignBERT effectively models hand gestures as visual tokens and BEST successfully applies BERT-style pre-training to triplet units, these methods predominantly focus on manual features, potentially neglecting crucial non-manual elements like facial expressions and body postures. Similarly, MASA's motion-aware autoencoder, though powerful, faces an information bottleneck from single-positive contrastive learning.

In contrast, our method demonstrates significant advantages in several respects. First, our method benefits from the action descriptions of various body parts related to sign lan-

guage in the text features, emphasizing various action features of the human body related to sign language. By incorporating these text features, our model can more comprehensively understand sign language actions, including not only hand gestures but also other important information such as facial expressions and body postures. This helps the model capture subtle differences in sign language more accurately, achieving performance comparable to or even better than SOTA methods on the SLR-500 dataset. Furthermore, our method performs well in feature representation at both the joint and bone levels, indicating that the model can effectively utilize feature information at different levels. Particularly in the four-stream setting in Table 2, our method achieved an accuracy of 97.1%, demonstrating its powerful capability to handle more complex sign language action sequences.

Table 2. Performance comparison on the SLR-500 dataset.

Method	Accuracy (%)
ST-GCN [34]	90.0
SignBERT [19]	94.5
BEST [7]	95.4
MASA [8]	96.3
SSRL [35]	96.9
Ours (joint\joint_motion)	96.01\95.28
Ours (bone\bone_motion)	94.37\94.19
Ours (4 streams fusion)	97.1

4.2.2. Performance Comparison on AUTSL Dataset

To further verify the generalization capability of our method, we also conducted experiments on the AUTSL dataset. The AUTSL dataset is widely known for its inherent diversity and complexity, which poses even greater challenges for SLR models. As shown in Table 3, our method also demonstrated its excellent performance on the AUTSL dataset, achieving the highest Top-1 accuracy of 97.07% and Top-5 accuracy of 99.89%, representing a significant improvement over previous state-of-the-art methods.

Table 3. Performance comparison on AUTSL dataset (Top-1 and Top-5 accuracy in %).

Method	Top-1	Top-5
SL-TSSI-DenseNet [36]	93.13	–
SSTCN [37]	93.37	–
SAM-SLR-V1 [37]	95.45	99.25
AM-GCN-A [38]	96.27	99.48
SAM-SLR-V2 [39]	96.47	99.76
TMS-Net [40]	96.62	99.71
SML [41]	96.85	99.79
Ours	97.07	99.89

Performance improvement can be attributed to the multi-positive contrastive learning mechanism, which provides a boost in accuracy of 1.36% by capturing inter-sign variations. Furthermore, our multimodal fusion strategy outperforms existing methods through the comprehensive integration of manual and non-manual features. These results confirm our method's strong generalization across different sign languages and datasets, maintaining consistent performance advantages while handling AUTSL's inherent variability challenges.

4.2.3. Comprehensive Performance Analysis

The evaluation was conducted on two well-balanced datasets. The SLR500 dataset comprises 500 classes, each containing 2500 samples. The AUTSL dataset consists of

226 classes with an average of 169.6 samples per class. To provide a thorough assessment, we supplement the accuracy metric with macro-averaged precision, recall, and F1-score, as presented in Table 4.

The tight clustering of these metrics around the accuracy values ($\Delta < 0.6\%$ across all metrics) confirms consistent performance across all sign categories. This metric alignment demonstrates that the high accuracy reliably reflects uniform classification capability without class-specific biases. The Top-5 accuracy further indicates robust sign identification, with correct predictions consistently appearing in the top results.

Table 4. Comprehensive evaluation metrics on balanced SLR500 and AUTSL datasets.

Dataset	Accuracy	Precision	Recall	F1-Score	Top-5 Acc
SLR500	97.1%	96.5%	96.8%	96.7%	99.6%
AUTSL	97.07%	96.2%	96.8%	96.5%	99.89%

4.3. Ablation Study

4.3.1. Combining Sign Language with LLMs

To validate the semantic quality of our generated descriptions, we conducted a detailed ablation study, with results presented in Table 5. This study indirectly assesses description quality by measuring its impact on the final recognition accuracy. The study analyzes the impact of using an unguided LLM, an expert-validated Knowledge Base (KB) alone, and our full RAG-based system against a Visual Encoder (VE)-only baseline.

Table 5. Ablation study on the description generation components. VE: Visual Encoder, KB: Knowledge Base.

VE	LLM	KB	Optimized Prompt	Accuracy (%)
✓	–	–	–	93.85
✓	✓	–	–	93.57
✓	–	✓	–	94.50
✓	✓	✓	–	94.89
✓	✓	✓	✓	95.25

✓ Indicates that the component was included in the experimental setup.

Our analysis begins with a visual-only baseline, which achieves 93.85% accuracy. When we introduce descriptions generated by an unguided LLM ("+LLM only"), performance degrades by 0.28%. This confirms that general-purpose LLMs are prone to generating semantically incorrect or "hallucinated" descriptions for a specialized domain like sign language, which misleads the model.

In contrast, when we augment the Visual Encoder with text solely from our expert-validated Knowledge Base ("+KB only"), accuracy increases by 0.65%. This result establishes a crucial baseline: the expert-sourced text is semantically sound and provides valuable supervisory signals.

Finally, our full system employs Retrieval-Augmented Generation (RAG) to ground the LLM in the expert KB. This approach achieves the best performance, with a significant +1.40% improvement over the baseline. This demonstrates a powerful synergy: the KB ensures the factual correctness and semantic relevance of the generated text, effectively eliminating hallucinations, while the LLM provides linguistic diversity and rephrasing that enriches the descriptions beyond the static KB entries. The substantial performance gain serves as strong evidence that our full system generates high-quality, semantically correct, and varied descriptions that are highly beneficial for the recognition task.

4.3.2. Multipart Contrastive Learning for Sign Language Recognition

Our method enhances traditional classification-based SLR by introducing a novel multipart contrastive learning mechanism. This approach is designed to effectively leverage LLM-generated knowledge (detailed action descriptions, synonyms, and expert-grounded part-specific texts), thereby fostering a deeper understanding of sign language semantics by establishing a stronger alignment between visual and textual representations.

The experimental results in Table 6 show a successive improvement in performance with each additional element introduced. The inclusion of synonym variants contributes an increase of +0.65%, while the use of optimized prompts adds another +0.88% to the accuracy. Moreover, incorporating precise part-specific descriptions delivers the most substantial improvement of +1.40%. When all these elements are fully integrated into the complete method, it achieves an impressive accuracy of 95.81%. This validates the effectiveness of our contrastive learning strategy, which successfully bridges the gap between visual and textual representation spaces, captures the nuanced differences in actions, and combines expert knowledge with the capabilities of language models.

Table 6. Ablation study of hierarchical multipart contrastive learning (based on joint data).

VE	Synonym	Prompt	Multipart	Accuracy (%)
✓	–	–	–	93.85
✓	✓	–	–	94.50
✓	–	✓	–	94.93
✓	–	–	✓	95.25
✓	✓	✓	–	95.38
✓	✓	✓	✓	95.81

✓ Indicates the inclusion of the corresponding element or knowledge source.

4.3.3. Keypoint Selection Analysis

The selection of keypoint combinations significantly impacts the recognition performance in skeleton-based SLR. Table 7 compares various approaches, demonstrating that our method achieves superior accuracy (93.85%), with 87 keypoints, outperforming existing combinations (27–76 keypoints) while maintaining computational efficiency.

Table 7. Comparison of the effects of different keypoint combinations in SLR.

Method	Num. Keypoints	Acc (%)	Parts	+Multipart (%)
all	133	59.22	–	–
base	27	93.46	3	93.62 (0.16%↑)
MASA	49	93.66	3	94.01 (0.35%↑)
Cosign	76	93.57	5	94.52 (0.95%↑)
Ours	87	93.85	5	95.25(1.40%↑)

There are limitations to previous keypoint selection strategies. SAM-SLR [37] significantly improved recognition accuracy by reducing the number of keypoints from 133 to 27. However, these methods have certain limitations in keypoint selection. SAM-SLR and MASA [8] neglect facial, lip, and other non-hand information related to sign language; although CoSign [6] collects more keypoints, it still does not cover all relevant body parts.

Through systematic examination of expert sign language descriptions, we identify 87 keypoints covering both manual (hands) and non-manual (face, mouth, body) locations commonly used in real-world signing. When combined with our multipart contrastive learning, this selection achieves a 1.4% accuracy improvement, which is significantly higher

than the gains of other methods (0.16–0.95%). The results validate that complete coverage and textual grounding yield optimal recognition performance.

4.3.4. Analysis of the Weighting Parameter α

The hyperparameter α in our composite loss function, $\mathcal{L}_{total} = \mathcal{L}_{cls} + \alpha \cdot \mathcal{L}_{con}^{multi}$, balances the influence of the standard classification loss ($\mathcal{L}_{cls}$) and our proposed multi-positive contrastive loss ($\mathcal{L}_{con}^{multi}$). To determine its optimal value, we conducted a comprehensive sensitivity analysis on the SLR-500 dataset, with results presented in Figure 5.

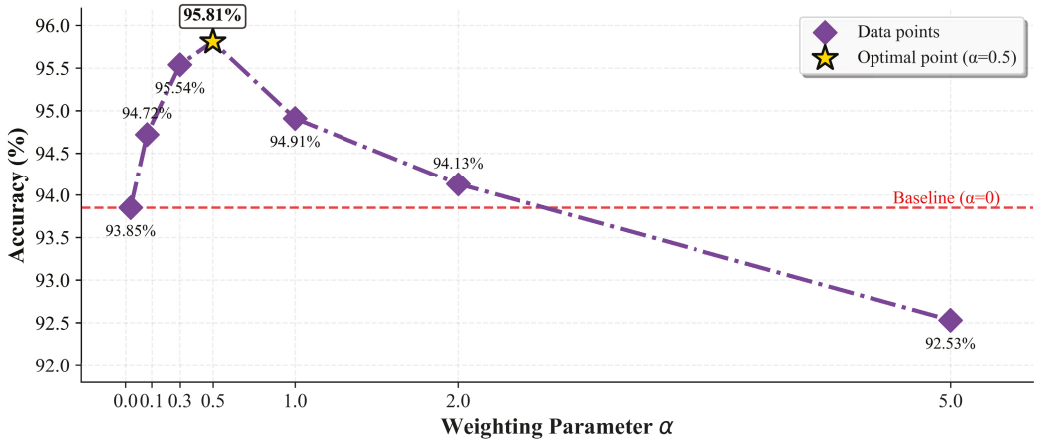


Figure 5. Sensitivity analysis of the weighting parameter α on the SLR-500 dataset. The baseline ($\alpha = 0$) uses only the classification loss.

The baseline model, using only $\mathcal{L}_{cls}$ ($\alpha = 0$), achieved 93.85 % accuracy. As α increases from 0 to 0.5, the contrastive loss provides effective regularization, guiding the model to learn more semantically rich features, which leads to a steady performance improvement. The accuracy peaks at 95.81% when $\alpha = 0.5$, indicating that a balanced contribution from both loss components is optimal. However, when α increases beyond 0.5, the total loss becomes overly dominated by the contrastive objective, causing the final accuracy to decline. Notably, at a very large value like $\alpha = 5.0$, the performance drops to 92.53 %, even below the baseline, confirming that an excessive focus on the auxiliary contrastive task is detrimental. This analysis validates our empirical choice of $\alpha = 0.5$ for all main experiments.

4.3.5. Analysis of Multipart Feature Contribution

To explicitly address our method's effectiveness in capturing non-manual features (e.g., face, mouth, body), as mentioned by prior work, we conducted an ablation study on the SLR-500 dataset to quantify the contribution of different anatomical parts. We systematically removed feature groups corresponding to the face, hands, and body from our full model (VE + Multipart) and measured the impact on accuracy. The results are presented in Table 8.

Table 8. Ablation study on the contribution of different body part features. The baseline is the Visual Encoder (VE) only, without multipart contrastive learning.

Configuration	Accuracy (%)	Performance Drop
VE-only (Baseline)	93.85	-
VE + Multipart (Full Model)	**95.25**	-
w/o Face features	94.55	−0.70
w/o Body features	95.04	−0.21
w/o Hands features	93.92	−1.33

The results quantify the importance of each component. As expected, hand features are the most critical, as their removal causes the largest performance drop (−1.33%). However, the experiments also provide strong quantitative evidence for the contribution of non-manual features. The removal of facial features leads to a significant decrease in accuracy of 0.70%, while the removal of body posture features results in a 0.21% drop. This confirms that our model successfully captures and leverages cues from facial expressions and body movements, moving beyond the limitations of methods that predominantly focus on hands.

Furthermore, the qualitative analysis of attention maps in Figure 6 corroborates these findings, showing that our model learns to focus on relevant facial or body regions for specific signs, unlike baselines that might over-attend to hands. Collectively, this analysis validates that our multipart framework effectively integrates both manual and non-manual information for more robust and accurate sign language recognition.



Figure 6. Attention Visualization Comparison: Skeleton-only (**Top**) baseline showing diffuse attention patterns. Our method's refined attention (**Bottom**) focusing on semantically relevant joints and relationships. Color coding: pink nodes = joint attention (size indicates importance), blue links = joint relationships (width/opacity indicate strength).

4.4. Visualized Analysis

To provide qualitative and deeper insights into our method's behavior, we conduct a comprehensive visual analysis using the SLR-500 trained model. This examination reveals how the integration of textual descriptions influences both spatial attention patterns and categorical performance across diverse sign categories.

4.4.1. Visualization of Human Keypoint Attention Weights

In Figure 6, we visualize the attention patterns of our skeleton encoder to analyze the impact of contrastive learning based on multipart descriptions. The attention weights are derived from channel-averaged feature maps, where joint importance is computed through normalized attention aggregation. Before incorporating text features, the model exhibits diffuse attention across multiple joints, indicating weaker semantic alignment. After applying our MC method with text guidance, the model significantly reduces attention noise, focusing more precisely on linguistically meaningful regions. For instance, in the Class "Tongue", the model correctly emphasizes left-hand and facial joints, while the Class "Ear"

shows enhanced attention to relevant inter-joint relationships. The visualization confirms that our method not only suppresses irrelevant keypoints but also strengthens the model's ability to capture sign language semantics.

4.4.2. Class-Wise Performance Gains

Figure 7 presents a detailed analysis of the class-wise accuracy improvements of our GSP-MC method versus the baseline CTR-GCN model on SLR-500. The visual-text contrastive learning method produces markedly different impacts across sign categories, revealing important patterns about its operational mechanisms.

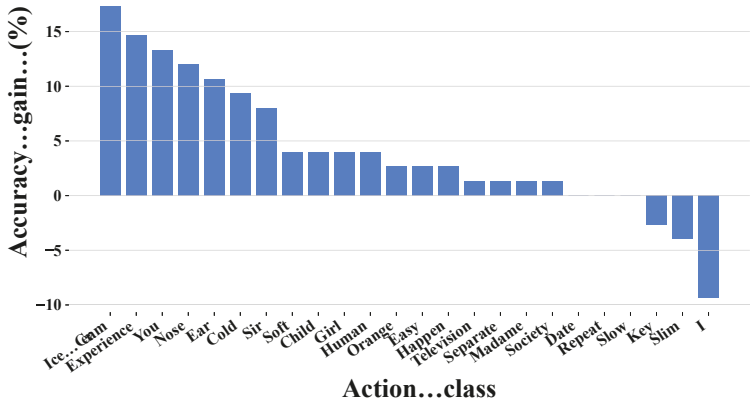


Figure 7. Class-wise Performance Gains.

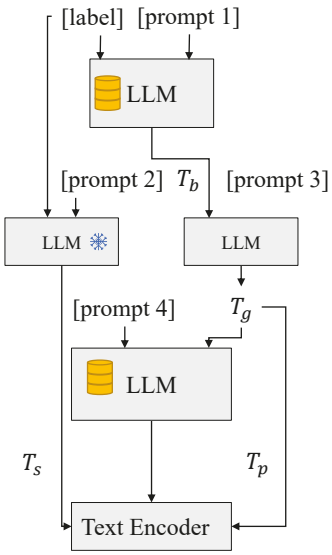
The substantial 17.33% performance gain for "ice cream" stems from its precise description specifying "bend fingers + press cheek + pinch thumb-index", while "experience" improves by a notable 14.67% with explicit "circular motion near forehead" trajectory details. Conversely, "thick" declines by 12% due to its vague "press down" description lacking critical body part and trajectory specifics, and "I" drops by 9.33% when generated texts confuse pronouns. Recognition accuracy proves to be directly dependent on complete descriptions, particularly when specifying exact body configurations ("bend fingers"), anatomical locations ("cheek"), and motions ("circular").

These results collectively demonstrate that our approach excels most for signs with concrete physical referents that are amenable to clear part-specific decomposition. Performance variations, both positive and negative, directly correlate with description quality. The fact that 4/5 of the classes maintained or exceeded baseline performance confirms the overall robustness of the multipart contrastive learning approach.

4.4.3. Case Study: Illustrating the Many-to-Many Mapping

To illustrate the many-to-many mapping mechanism, which is central to our multi-positive contrastive learning, we present a case study of the Chinese sign "Postman" from the SLR-500 dataset. As shown in Figure 8, our GSP module generates a hierarchical set of textual descriptions for this single sign instance. This process exemplifies the one-sign-to-many-texts relationship, where a single skeleton sequence is associated with multiple positive text anchors. Specifically, these anchors include (1) a refined global description T_g, such as "(1) Gently touch the tip of the thumb and middle finger on the palm of the other hand, mimicking the action of sticking a stamp. (2) Extend one hand forward, as if handing something over. (3) Perform the "driver" gesture."; (2) a collection of synonym variants T_s, such as "mail carrier" and "postal worker"; and (3) a set of atomic part-level descriptions T_p segmented from T_g.

Furthermore, the part-level descriptions establish an explicit many-texts-to-many-skeleton-parts alignment. For instance, a single text segment like T_{p1} ("...on the palm of the other hand.") describes a coordinated action involving both hands and is therefore mapped to both the left-hand S_{lh} and right-hand S_{rh} skeleton feature groups. Conversely, a single skeletal component, such as the right-hand group S_rh, participates in multiple distinct actions described by different text segments T_{p1}, T_{p2}, and T_{p3}. This example illustrates how our method moves beyond a simple one-to-one paradigm. By simultaneously optimizing alignments between a single sign's skeleton data and its rich, multi-faceted textual descriptions, the model learns a feature representation that is robust to variations and grounded in the compositional nature of sign language.



Label: Postman

$\boldsymbol{T_b}$: (1) Gently touch the tip of the thumb and middle finger on the palm of the other hand, mimicking the action of sticking a stamp. (2) Extend one hand forward, as if handing something over. (3) Perform the "driver" gesture.

$\boldsymbol{T_s}$: mail carrier, postman, postal worker, letter carrier, courier, mailman, delivery driver, parcel delivery person, postwoman, postie

$\boldsymbol{T_g}$: (1) Gently touch the tip of the thumb and middle finger on the palm of the other hand. (2) Extend one hand forward. (3) Thumb and index finger form a circle, place on the chest.

$\boldsymbol{T_p}$: Touch the tip of the thumb and middle finger on the palm of the other hand. 1,2; Extend hand forward, 2; Thumb and index finger make a small circle, place on the chest. 2,4;

Figure 8. An illustration of the many-to-many mapping generated by our GSP module for the Chinese sign "Postman". For a single sign instance, the module produces (a) a detailed global description (T_g), (b) a set of synonyms (T_s), and (c) segmented part-level descriptions (T_p). The numeric labels indicate the associated body part: 0 (unrelated), 1 (left hand), 2 (right hand), 3 (face), and 4 (body).

4.5. Efficiency and Computational Cost Analysis

To evaluate the practical viability and deployment potential of our method, we conducted a comprehensive analysis of its computational costs, comparing it against the CTR-GCN baseline. The results, including training time, GPU memory usage, inference speed, and model parameters, are summarized in Table 9.

Table 9. Computational cost analysis. Our method is compared with the baseline on an NVIDIA A100 GPU. The costs for our method are presented separately for the training and inference stages to provide a clear comparison.

Method	Training Time (h)	GPU Memory (GB)	Inference (ms/Sample)	Parameters (M)
CTR-GCN (Baseline)	8.2	12.4	4.2	5.6
Ours (Training)	21.3	24.6	-	69.1
Ours (Inference)	-	12.8	4.3	5.6

As shown in Table 9, the training phase of our method requires more resources than the baseline. The training time is approximately 2.6 times longer, and the GPU memory

usage is nearly doubled. This increased cost is attributable to three main factors: (1) the one-time, offline text generation and preprocessing using LLMs (3 h); (2) the additional forward passes required for the dual-encoder architecture; and (3) the computational demands of the multi-positive contrastive loss.

However, our central claim of efficiency pertains to the inference stage. At an average of 4.3 ms per sample, our model's inference speed translates to a throughput of approximately 232 samples per second (fps). This high performance surpasses typical real-time requirements (e.g., 30 or 60 fps) and confirms that our method introduces no meaningful overhead compared to the baseline (4.2 ms). This efficiency is possible because the text encoder (63.5 M parameters) is discarded after training, leaving the deployed model with the same parameter count (5.6 M) as the baseline. Moreover, since the inference time scales linearly with sequence length, our model robustly maintains its suitability for practical, real-time applications.

5. Discussion

While our method demonstrates strong performance, a comprehensive discussion of its limitations, practical significance, and deployment challenges is warranted to contextualize the results and guide future research.

5.1. Limitations and Future Directions

Our study has several limitations that open avenues for future work:

1. The most significant limitation is our evaluation on isolated sign datasets. While this provides a controlled environment to validate our core approach, real-world communication involves continuous signing. Extending our method to Continuous Sign Language Recognition (CSLR) presents substantial challenges, including (a) temporal segmentation to identify sign boundaries; (b) modeling co-articulation effects, where the appearance of a sign is influenced by adjacent signs; and (c) interpreting grammatical structures conveyed over time. Future work will focus on adapting our framework for CSLR, potentially by integrating it into a sequence-to-sequence architecture (e.g., Transformer or CTC-based models) to process continuous streams on datasets like RWTH-PHOENIX-Weather [42] and CSL-Daily [43].
2. Our method's performance is contingent on the accuracy of the upstream pose estimation model. In real-world scenarios with challenging conditions such as poor lighting, motion blur, or occlusions, the quality of the extracted skeleton data may degrade, consequently affecting recognition accuracy. Improving robustness to imperfect pose estimation is a key direction for future research.
3. The LLM-based description generation introduces a one-time computational cost for each vocabulary item. While this is an offline process and does not affect inference speed, it adds complexity to the training pipeline compared to traditional methods.
4. Although we have demonstrated success on both Chinese (CSL) and Turkish (AUTSL) sign languages, our model has not been tested on a wider range of sign languages or the vast stylistic and regional variations that exist within them. Future work should explore the model's generalizability and potential need for fine-tuning on diverse linguistic and signer-specific data.

5.2. Practical Significance and Deployment Considerations

Despite these limitations, our method has significant practical implications:

1. The state-of-the-art accuracy (>97%) on large-scale isolated sign datasets makes the method suitable for immediate deployment in applications where single-sign interac-

tion is prevalent. This includes educational software for learning sign language, assistive communication boards, and command-based control systems for smart devices.
2. By leveraging fine-grained textual descriptions, our model learns a more semantically rich representation of signs compared to methods relying solely on class labels. This could enable more nuanced applications, such as sign-to-text systems that can differentiate between visually similar but semantically distinct signs.
3. Importantly, the LLM and text encoder are only used during the training phase. During inference, our model operates solely on skeleton data, maintaining the same computational efficiency as standard skeleton-based classifiers. This makes it practical for real-time deployment on resource-constrained devices.

However, real-world deployment must address several challenges: (a) ensuring the entire pipeline, from video capture to recognition, meets real-time processing requirements; (b) building robustness to diverse signing styles, speeds, and environmental conditions; (c) ensuring seamless integration with existing accessibility infrastructure; and (d) navigating data privacy and ethical considerations related to the collection and use of sign language data.

6. Conclusions

This paper introduced GSP-MC, which achieves significant performance gains through three key theoretical innovations. First, the Generative Sign-description Prompts succeed where they address the semantic gap between visual features and linguistic meaning, using RAG to ground LLM outputs in expert knowledge, prevent hallucinations, and maintain descriptive richness. This provides a 0.88% improvement over raw LLM integration, demonstrating that domain-specific knowledge anchoring is essential for specialized recognition tasks.

Second, the multi-positive contrastive learning proves effective by acknowledging the inherent many-to-many relationship in sign language: one sign has multiple valid executions and descriptions. Unlike single-positive methods that force artificial one-to-one mappings, our probabilistic alignment (Equation (9)) naturally handles variation, contributing 1.40% accuracy gain. This theoretical shift from deterministic to probabilistic matching better reflects the linguistic reality of sign language.

Third, the hierarchical part-based decomposition works because sign language is compositionally structured, where meaning emerges from coordinated movements of hands, face, and body. By explicitly modeling these components (87 keypoints across five body parts), we capture both local dynamics and global semantics. Ablation studies confirm that this design choice yields the largest individual improvement (1.40%), validating that architectural alignment with linguistic structure enhances recognition.

The method's cross-lingual success (97.1% Chinese, 97.07% Turkish) demonstrates that these principles transcend specific sign languages. The key insight is that bridging vision and language requires not just technical integration but theoretical alignment with how signs convey meaning through structured, multi-component expressions with natural variation. Future work should explore these principles in continuous signing and sign language translation tasks.

Supplementary Materials: The following supporting information can be downloaded at: https://github.com/LiangSiyv/GSPMC (accessed on 21 September 2025).

Author Contributions: Conceptualization, S.L.; methodology, S.L.; validation, H.C.; formal analysis, K.L.; investigation, W.X.; resources, W.X.; writing, original draft preparation, S.L.; writing, review and editing, K.L. and Y.L.; visualization, X.L.; supervision, Q.M.; funding acquisition, Q.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Science and Technology Major Project, grant number 2022ZD0117103; the National Natural Science Foundation of China, grant number 62272364; the Provincial Key Research and Development Program of Shaanxi, grant number 2024GH-ZDXM-47; and the Research Project on Higher Education Teaching Reform of Shaanxi Province, grant number 23JG003.

Data Availability Statement: The original SLR-500 and AUTSL datasets are publicly available from their respective sources (SLR-500: https://ustc-slr.github.io/datasets/2015_csl/ (accessed on 21 September 2025), AUTSL: https://cvml.ankara.edu.tr/datasets/ (accessed on 21 September 2025)). Our Supplementary Materials will be hosted at https://github.com/LiangSiyv/GSPMC (accessed on 21 September 2025).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Microsoft. Developing with Kinect. 2017. Available online: https://developer.microsoft.com/en-us/windows/kinect/develop (accessed on 19 August 2025).
2. Intel. Intel® RealSense™ Technology. 2019. Available online: https://www.intel.com/content/www/us/en/architecture-and-technology/realsense-overview.html (accessed on 19 August 2025).
3. Fang, H.S.; Li, J.; Tang, H.; Xu, C.; Zhu, H.; Xiu, Y.; Li, Y.L.; Lu, C. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 7157–7173. [CrossRef] [PubMed]
4. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep high-resolution representation learning for human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5693–5703. [CrossRef]
5. Lin, K.; Wang, X.; Zhu, L.; Zhang, B.; Yang, Y. SKIM: Skeleton-based isolated sign language recognition with part mixing. *IEEE Trans. Multimed.* **2024**, *26*, 4271–4280. [CrossRef]
6. Jiao, P.; Min, Y.; Li, Y.; Wang, X.; Lei, L.; Chen, X. Cosign: Exploring co-occurrence signals in skeleton-based continuous sign language recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 20676–20686. [CrossRef]
7. Zhao, W.; Hu, H.; Zhou, W.; Shi, J.; Li, H. BEST: BERT pre-training for sign language recognition with coupling tokenization. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; Volume 37, pp. 3597–3605. [CrossRef]
8. Zhao, W.; Hu, H.; Zhou, W.; Mao, Y.; Wang, M.; Li, H. MASA: Motion-aware Masked Autoencoder with Semantic Alignment for Sign Language Recognition. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 10793–10804. [CrossRef]
9. Liang, S.; Li, Y.; Shi, Y.; Chen, H.; Miao, Q. Integrated multi-local and global dynamic perception structure for sign language recognition. *Pattern Anal. Appl.* **2025**, *28*, 50. [CrossRef]
10. Yuan, L.; He, Z.; Wang, Q.; Xu, L.; Ma, X. Skeletonclip: Recognizing skeleton-based human actions with text prompts. In Proceedings of the 2022 8th International Conference on Systems and Informatics (ICSAI), Kunming, China, 10–12 December 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1–6. [CrossRef]
11. Bao, H.; Dong, L.; Wei, F.; Wang, W.; Yang, N.; Liu, X.; Wang, Y.; Gao, J.; Piao, S.; Zhou, M.; et al. UniLMv2: Pseudo-Masked Language Models for Unified Language Model Pre-Training. In Proceedings of the 37th International Conference on Machine Learning, Virtual, 13–18 July 2020; Proceedings of Machine Learning Research; Volume 119, pp. 642–652.
12. Liu, R.; Liu, Y.; Wu, M.; Xin, W.; Miao, Q.; Liu, X.; Li, L. SG-CLR: Semantic representation-guided contrastive learning for self-supervised skeleton-based action recognition. *Pattern Recognit.* **2025**, *162*, 111377. [CrossRef]
13. Xiang, W.; Li, C.; Zhou, Y.; Wang, B.; Zhang, L. Generative Action Description Prompts for Skeleton-based Action Recognition. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 10276–10285. [CrossRef]
14. Hwang, E.J.; Cho, S.; Lee, J.; Park, J.C. An Efficient Sign Language Translation Using Spatial Configuration and Motion Dynamics with LLMs. *arXiv* **2024**, arXiv:2408.10593 . [CrossRef]
15. Yan, T.; Zeng, W.; Xiao, Y.; Tong, X.; Tan, B.; Fang, Z.; Cao, Z.; Zhou, J.T. Crossglg: Llm guides one-shot skeleton-based 3d action recognition in a cross-level manner. In Proceedings of the European Conference on Computer Vision, Milan, Italy, 29 September–4 October 2024; Springer: Berlin/Heidelberg, Germany, 2024; Volume 15078, pp. 113–131. [CrossRef]
16. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R.B. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Computer Vision Foundation, Seattle, WA, USA, 13–19 June 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 9729–9738. [CrossRef]

17. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. In Proceedings of the 37th International Conference on Machine Learning, Online, 13–18 July 2020; ICML 2020, Virtual Event, Proceedings of Machine Learning Research; Volume 119, pp. 1597–1607. Available online: http://proceedings.mlr.press/v119/chen20j.html (accessed on 21 September 2025).
18. Liu, K.; Ma, Z.; Kang, X.; Li, Y.; Xie, K.; Jiao, Z.; Miao, Q. Enhanced Contrastive Learning with Multi-view Longitudinal Data for Chest X-ray Report Generation. *arXiv* **2025**, arXiv:2502.20056. [CrossRef]
19. Hu, H.; Zhao, W.; Zhou, W.; Wang, Y.; Li, H. SignBERT: Pre-Training of Hand-Model-Aware Representation for Sign Language Recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 11087–11096. [CrossRef]
20. Zuo, R.; Wei, F.; Mak, B. Natural Language-Assisted Sign Language Recognition. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 14890–14900. [CrossRef]
21. Zhang, H.; Guo, Z.; Yang, Y.; Liu, X.; Hu, D. C2st: Cross-modal contextualized sequence transduction for continuous sign language recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 21053–21062. [CrossRef]
22. Kalakonda, S.S.; Maheshwari, S.; Sarvadevabhatla, R.K. Action-gpt: Leveraging large-scale language models for improved and generalized action generation. In Proceedings of the 2023 IEEE International Conference on Multimedia and Expo (ICME), Brisbane, Australia, 10–14 July 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 31–36. [CrossRef]
23. Wong, R.; Camgöz, N.C.; Bowden, R. Sign2GPT: Leveraging Large Language Models for Gloss-Free Sign Language Translation. In Proceedings of the The Twelfth International Conference on Learning Representations, Vienna, Austria, 7–11 May 2024.
24. Chen, Y.; Zhang, Z.; Yuan, C.; Li, B.; Deng, Y.; Hu, W. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 13359–13368. [CrossRef]
25. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 8748–8763. Available online: http://proceedings.mlr.press/v139/radford21a.html (accessed on 21 September 2025).
26. Oord, A.v.d.; Li, Y.; Vinyals, O. Representation learning with contrastive predictive coding. *arXiv* **2018**, arxiv:1807.03748. Available online: http://arxiv.org/abs/1807.03748 (accessed on 21 September 2025).
27. Tian, Y.; Fan, L.; Isola, P.; Chang, H.; Krishnan, D. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 48382–48402. Available online: http://papers.nips.cc/paper_files/paper/2023/hash/971f1e59cd956cc094da4e2f78c6ea7c-Abstract-Conference.html (accessed on 21 September 2025).
28. Huang, J.; Zhou, W.; Li, H.; Li, W. Attention-based 3D-CNNs for large-vocabulary sign language recognition. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *29*, 2822–2832. [CrossRef]
29. Sincan, O.M.; Keles, H.Y. Autsl: A large scale multi-modal turkish sign language dataset and baseline methods. *IEEE Access* **2020**, *8*, 181340–181355. [CrossRef]
30. Gu, D.; Wei, D.; Yang, Y.;Wang, C.; Yu, Y.; Gao, H.; Wu, Y.; Heng, M.; Qiu, B.; Liu, W.; et al. *Chinese Sign Language (Revised Edition) (Parts 1 & 2)*; Huaxia Publishing House: Beijing, China, 2018.
31. Dingqian, G.; Dan, W.; Yang, Y.; Chenhua, W.; Yuanyuan, Y.; Hui, G.; Yongsheng, W.; Miao, H.; Bing, Q.; Wa, L.; et al. *Lexicon of Common Expressions in Chinese National Sign Language*; Technical Report GF 0020–2018; National Standard; Ministry of Education of the People's Republic of China, State Language Commission, China Disabled Persons' Federation: Beijing, China, 2018.
32. Dingqian, G.; Dan, W.; Chenhua, W.; Hui, G.; Yuanyuan, Y.; Miao, H.; Bing, Q.; Yongsheng, W. *Chinese Manual Alphabet*; Technical Report GF 0021–2019; National Standard; Ministry of Education of the People's Republic of China, State Language Commission, China Disabled Persons' Federation: Beijing, China, 2019.
33. Kavak, S. *Turkish Sign Language Dictionary*; National Education Ministry: New Delhi, India, 2015.
34. Yan, S.; Xiong, Y.; Lin, D. Spatial temporal graph convolutional networks for skeleton-based action recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, IL, USA, 2–7 February 2018; Volume 32. [CrossRef]
35. Zhao, W.; Zhou, W.; Hu, H.; Wang, M.; Li, H. Self-supervised representation learning with spatial-temporal consistency for sign language recognition. *IEEE Trans. Image Process.* **2024**, *33*, 4188–4201. [CrossRef] [PubMed]
36. Laines, D.; Gonzalez-Mendoza, M.; Ochoa-Ruiz, G.; Bejarano, G. Isolated sign language recognition based on tree structure skeleton images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 276–284. [CrossRef]
37. Jiang, S.; Sun, B.; Wang, L.; Bai, Y.; Li, K.; Fu, Y. Sign language recognition via skeleton-aware multi-model ensemble. *arXiv* **2021**, arXiv:2110.06161. Available online: https://arxiv.org/abs/2110.06161 (accessed on 21 September 2025).

38. Liu, Y.; Lu, F.; Cheng, X.; Yuan, Y. Asymmetric multi-branch GCN for skeleton-based sign language recognition. *Multimed. Tools Appl.* **2024**, *83*, 75293–75319. [CrossRef]
39. Jiang, S.; Sun, B.; Wang, L.; Bai, Y.; Li, K.; Fu, Y. Skeleton aware multi-modal sign language recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Computer Vision Foundation, Nashville, TN, USA, 19–25 June 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 3413–3423. [CrossRef]
40. Deng, Z.; Leng, Y.; Chen, J.; Yu, X.; Zhang, Y.; Gao, Q. TMS-Net: A multi-feature multi-stream multi-level information sharing network for skeleton-based sign language recognition. *Neurocomputing* **2024**, *572*, 127194. [CrossRef]
41. Deng, Z.; Leng, Y.; Hu, J.; Lin, Z.; Li, X.; Gao, Q. SML: A Skeleton-based multi-feature learning method for sign language recognition. *Knowl.-Based Syst.* **2024**, *301*, 112288. [CrossRef]
42. Forster, J.; Schmidt, C.; Hoyoux, T.; Koller, O.; Zelle, U.; Piater, J.H.; Ney, H. RWTH-PHOENIX-Weather: A Large Vocabulary Sign Language Recognition and Translation Corpus. In Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, 23–25 May 2012; Calzolari, N., Choukri, K., Declerck, T., Dogan, M.U., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Eds.; European Language Resources Association (ELRA): Paris, France, 2012; pp. 3785–3789.
43. Zhou, H.; Zhou, W.; Qi, W.; Pu, J.; Li, H. Improving Sign Language Translation with Monolingual Data by Sign Back-Translation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, Nashville, TN, USA, 19–25 June 2021; Computer Vision Foundation; IEEE: Piscataway, NJ, USA, 2021; pp. 1316–1325. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

sensors

MDPI

Article

Image Sensor-Supported Multimodal Attention Modeling for Educational Intelligence

Yanlin Chen [1,2,†], Yingqiu Yang [1,2,†], Zeyu Lan [1,†], Xinyuan Chen [1], Haoyuan Zhan [1], Lingxi Yu [1] and Yan Zhan [1,3,*]

[1] National School of Development, Peking University, Beijing 100871, China
[2] College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China
[3] Artificial Intelligence Research Institute, Tsinghua University, Beijing 100084, China
* Correspondence: yan_zhan_thu@163.com
† These authors contributed equally to this work.

Abstract

To address the limitations of low fusion efficiency and insufficient personalization in multimodal perception for educational intelligence, a novel deep learning framework is proposed that integrates image sensor data with textual and contextual information through a cross-modal attention mechanism. The architecture employs a cross-modal alignment module to achieve fine-grained semantic correspondence between visual features captured by image sensors and associated textual elements, followed by a personalized feedback generator that incorporates learner background and task context embeddings to produce adaptive educational guidance. A cognitive weakness highlighter is introduced to enhance the discriminability of task-relevant features, enabling explicit localization and interpretation of conceptual gaps. Experiments show the proposed method outperforms conventional fusion and unimodal baselines with 92.37% accuracy, 91.28% recall, and 90.84% precision. Cross-task and noise-robustness tests confirm its stability, while ablation studies highlight the fusion module's +4.2% accuracy gain and the attention mechanism's +3.8% recall and +3.5% precision improvements. These results establish the proposed method as a transferable, high-performance solution for next-generation adaptive learning systems, offering precise, explainable, and context-aware feedback grounded in advanced multimodal perception modeling.

Keywords: image sensors; visual-text integration; multimodal perception modeling; cross-modal alignment; educational intelligence

1. Introduction

With the rapid advancement of intelligent education, AI-driven efficient, personalized, and interpretable teaching feedback has become a key research focus in educational technology [1]. Traditional feedback relies heavily on manual teacher evaluation and experience, supplemented by inputs from students, peers, and supervisors, which is time-consuming and often inadequate for large-scale personalized teaching needs [2]. Advances in deep learning, especially large-scale models, enable the integration and analysis of multimodal data such as images and text, facilitating better understanding of teaching behaviors and intelligent feedback generation [3,4]. Li et al. demonstrated that intelligent tutoring systems can analyze student behaviors, emotions, progress, and answer accuracy in real time, delivering precise personalized feedback to support teaching strategy adjustments [5]. Combining image analysis with natural language processing allows examination

of non-verbal cues (e.g., facial expressions, posture) alongside verbal outputs, enhancing feedback comprehensiveness and accuracy [5]. The adoption of large-model AI technologies improves classroom efficiency and effectiveness, driving transformative changes in educational technology [6,7].

Most current teaching feedback systems primarily handle structured text inputs. Abdi et al. developed DTLP, an automated deep learning system analyzing student feedback with features including word embeddings, affective knowledge, emotion rules, statistics, and linguistic characteristics [8]. Shaik et al. integrated NLP methods to annotate and respond to student feedback, enhancing AI's educational impact [9]. However, these systems show limitations in jointly modeling student work images (handwritten assignments, lab reports, sketches) and textual descriptions (reasoning, reflections), restricting comprehensive understanding of learning outputs. The rise of multimedia data has increased interest in semantic alignment across modalities [10]. Despite progress, deeper linguistic understanding remains a challenge for generating effective feedback. Joint image–text modeling offers a holistic view of student work, enabling precise personalized feedback and addressing shortcomings of traditional approaches [11]. Deploying large language models (LLMs) in education also faces critical challenges. At the semantic level, LLMs have notable limitations in parsing complex linguistic phenomena: Weissweiler et al. [12] showed that mainstream models like GPT and Gemini fail to capture additional semantics in constructions such as causative motion (e.g., "sneeze causing object movement"), revealing gaps in deep semantic reasoning. These semantic weaknesses affect text generation and introduce linguistic patterns with privacy and security implications. Muñoz-Ortiz et al. [13] found LLM-generated news texts exhibit simpler syntax, emotional imbalance (overemphasis on "joy"), and overuse of pronouns, numerals, and auxiliaries—linguistic traits linked to privacy risks such as entity reference obfuscation and attribute inference [14].

Recent work suggests that injecting explicit linguistic knowledge can mitigate these issues. Zhang et al. [15] proposed LINGOLLM, which integrates grammar books and dictionaries into GPT-4, enabling efficient processing of endangered languages and achieving a BLEU score of 10.5, thus opening paths for low-resource language applications. This underscores the value of structured linguistic knowledge in enhancing multimodal and multilingual educational AI. Deep learning advancements in multimodal learning and joint image–text modeling have spurred interest in integrated analysis and feedback systems. Kumar et al. developed a system generating human-like feedback from paired text and image inputs [16]. Xie et al. highlighted multimodal interaction's advantages in education, combining speech, text, and visuals for richer outputs [17]. Liu et al. tackled personalized multimodal feedback generation, enabling customized responses across disciplines for assignments involving images, audio, and text [18]. To address these challenges, we propose a multimodal Transformer-based framework for teaching feedback generation, combining student-submitted images and textual descriptions with cross-modal attention. The primary contributions of this study are as follows:

1. A multimodal Transformer framework for automated teaching feedback generation is proposed, jointly modeling student-submitted images and textual content.
2. Three innovative modules are introduced: the image–text semantic alignment module, the personalized feedback generation module, and the cognitive weakness highlighter module.
3. Experimental validation demonstrates that the proposed method achieves high accuracy and strong practicality in various educational tasks, such as essay scoring, design feedback, and answer correction.

2. Related Work

2.1. Research on Educational Feedback Systems and AI-Assisted Teaching

Traditional educational feedback has been teacher-centered, relying on grading, classroom interaction, and performance evaluation, often dependent on teachers' subjective judgment, causing delays, inconsistency, and limited personalization [19,20]. AI has transformed feedback systems by automating tasks like essay scoring, grading, and behavior analysis. Steiss et al. highlighted AI tools such as ChatGPT that use NLP and machine learning for real-time feedback generation [20]. Yet challenges remain in producing fine-grained, context-aware feedback and in integrating AI across diverse teaching settings [21]. While AI improves efficiency, objectivity, automation, and personalization, issues like limited accuracy, poor generalization, and low interpretability persist. Liu et al. noted AI often neglects younger students' emotional needs and grade-specific alignment, causing skepticism of AI-only scoring [22]. In intelligent tutoring and personalized learning, AI can deliver high-quality, individualized feedback that supports cognitive and metacognitive skills and positive emotions, but concerns about privacy, bias, unreliable feedback, negative perceptions, limited capability, academic integrity, and lack of guidance remain [23].

2.2. The Application of Multimodal Deep Learning in Education

With the continuous advancement of deep learning technologies, the integration of vision and language has emerged as a key research direction in AI, particularly in educational image–text analysis. Multimodal deep learning models, such as ViLT, BLIP (Bootstrapping Language-Image Pre-training), and CLIP (Contrastive Language-Image Pre-training), combine visual information (images, videos) and linguistic information (text, speech) to enable more comprehensive and intelligent understanding and analysis. ViLT (Vision-and-Language Transformer) is a visual–language fusion model that converts images into sequences of visual tokens and processes them together with text tokens through a Transformer for cross-modal learning [24]. By leveraging ViLT, educational systems can extract and analyze relevant information from both images and text in instructional materials, thus enabling automated educational content generation and learning resource recommendation.

BLIP is an innovative multimodal pre-training framework designed to enhance joint language–vision modeling through self-supervised learning [25]. CLIP, proposed by OpenAI, is a vision–language fusion model trained using contrastive learning to simultaneously optimize image and text representations [26]. It enables automated image–text matching in educational platforms and can generate relevant answers or feedback based on either modality, thus improving interactivity. Despite these advancements, vision–language fusion models still face limitations when processing complex structured multimodal educational content.

3. Materials and Method

3.1. Data Collection

To address the practical requirements of research on open-ended expression and structured critical feedback in law and economics education, the dataset was constructed under multimodal conditions, with broader coverage and a larger sample size to enhance model generalization and robustness across diverse instructional tasks, as shown in Table 1. Data collection was conducted using undergraduate courses from three comprehensive universities, encompassing both core and elective subjects such as Principles of Civil Law, Public Finance, Legal Writing Practice, Principles of Economics, International Economic Law, and Public Finance Management. The image data were sourced from scanned copies of paper-based assignments, scanned classroom experiment and re-

search reports, student-drawn argumentation diagrams, economic statistical charts, case analysis flowcharts, and photographs of classroom blackboard writing. Paper-based submissions were collected by course instructors and digitized using high-resolution scanners at 600 dpi. During acquisition, edge cropping, color correction, and contrast enhancement were applied to ensure the preservation of essential details and structural features. Certain assignments were completed directly on tablets or graphic tablets and submitted in PDF or JPEG format, maintaining an original resolution of no less than 2480×3508 pixels to facilitate subsequent image segmentation and feature extraction.

Table 1. Sample distribution of dataset used in this paper.

Task Type	Sample Count	Image Count	Text Count
Case analysis	1927	1298	3103
Statute application	1311	325	2037
Economic chart interpretation	1193	1905	1592
Policy commentary	781	379	1017

Text data were obtained from written responses and analytical reports submitted by students for assignments and projects, covering content types such as judgment reasoning, case commentary, economic policy review, statistical result interpretation, solution proposals, and reflective summaries. The original file formats included Word documents, PDF files, and text exported from online assignment platforms. All text data were collected with the original formatting, paragraphing, and punctuation preserved, along with metadata including task type, submission time, and assignment background, to support contextual modeling and feature analysis. In law-related tasks, the text typically contained statutory citations, case names, and logical argumentation structures, whereas in economics-related tasks, it more frequently included formula derivations, data interpretations, and chart references. Upon completion of collection, each image and text entry was paired via a unique identifier, forming one-to-one multimodal records to ensure consistency in cross-modal analysis.

Annotation guidelines were formulated by three senior law and economics educators, each with more than ten years of teaching and grading experience. The annotation dimensions included logical structure completeness, legal or economic theory application, accuracy of expression and use of technical terminology, and personalized improvement suggestions with reflective prompts. Each sample was independently annotated by at least two experts, with cross-checking performed to ensure consistency, and in cases of disagreement, a third annotator acted as arbiter to guarantee the authority and uniformity of the final annotations. The resulting dataset comprises 5000 multimodal assignment samples, with case analysis tasks accounting for the largest proportion, followed by statute application and economic chart interpretation, while policy commentary tasks were fewer but demonstrated diverse viewpoints and prominent critical thinking features. These data provide a solid foundation for achieving precise feedback generation in multidisciplinary and multi-expression-form educational scenarios.

3.2. Dataset Construction

Given the structural complexity, domain-specific characteristics, and expressive diversity of multimodal assignment data in the fields of law and economics, a systematic preprocessing and augmentation pipeline was designed prior to model training. This pipeline comprised three main stages: image processing, text processing, and data augmentation.

3.2.1. Image Processing

In the image processing stage, variations in resolution and clarity among assignment images from different sources were mitigated while key informational regions were preserved. All scanned images and handwritten sketches were first standardized in resolution by scaling them to a predefined size (H, W):

$$I' = \mathcal{R}(I, H, W),$$

where I denotes the original image, $\mathcal{R}(\cdot)$ represents the bicubic interpolation resampling function, and H and W are the target image height and width, respectively. After standardization, handwriting, structural lines, and edge features were enhanced by applying a sharpening operator $\mathcal{S}(\cdot)$ followed by an edge enhancement filter $\mathcal{E}(\cdot)$, yielding:

$$I'' = \mathcal{E}(\mathcal{S}(I')).$$

To facilitate subsequent cross-modal alignment, a patchifying method was employed to divide the image into non-overlapping blocks of size $P \times P$. In combination with a diagram detection model $\mathcal{G}(\cdot)$, blocks containing key diagrams or structural graphics were automatically identified and retained, thereby reducing redundant background noise.

3.2.2. Text Processing

In the text processing stage, a hierarchical procedure was implemented to accommodate the high density of domain-specific terminology, statute references, and policy entities in law and economics assignments, ensuring structural regularity and semantic accuracy. Initially, a domain-specific lexicon was applied for terminology recognition and normalization:

$$T' = \mathcal{N}_t(T),$$

where T is the original text and $\mathcal{N}_t(\cdot)$ denotes the terminology normalization function based on the domain lexicon. Subsequently, a normalization parser $\mathcal{L}(\cdot)$ was used to encode statute references in a consistent format, while a policy entity recognition function $\mathcal{P}(\cdot)$ extracted named entities relevant to economic and legal policies, forming a structured element set:

$$E = \mathcal{L}(T') \cup \mathcal{P}(T').$$

Based on this, a citation cleaning function $\mathcal{C}(\cdot)$ was applied to remove extraneous symbols, standardize inconsistent citation formats, and eliminate irrelevant noise. Finally, a syntactic restoration operator $\mathcal{S}_p(\cdot)$ and a spell-checking function $\mathcal{F}(\cdot)$ were applied to reconstruct the original grammatical structure and reduce the impact of syntactic flaws on model comprehension, producing a structurally complete and semantically clear text sequence T'':

$$T'' = \mathcal{F}(\mathcal{S}_p(\mathcal{C}(T'))).$$

3.2.3. Data Augmentation

In the data augmentation stage, a feedback contrastive learning strategy was introduced to construct positive and negative feedback sample pairs (x^+, x^-) under the same task context, where the content was similar but the quality of reasoning or viewpoint differed significantly. Semantic discrimination was achieved by minimizing the embedding distance between positive and anchor samples while maximizing the distance to negative samples:

$$\mathcal{L}_{con} = -\log \frac{\exp(\text{sim}(f(x), f(x^+))/\tau)}{\exp(\text{sim}(f(x), f(x^+))/\tau) + \exp(\text{sim}(f(x), f(x^-))/\tau)}.$$

Here, $f(\cdot)$ is the feature encoding function, $\text{sim}(\cdot,\cdot)$ denotes cosine similarity, and τ is the temperature parameter. Furthermore, a cross-task synonymous structure perturbation strategy $\mathcal{T}(\cdot)$ was developed, introducing structurally equivalent but lexically varied replacements across different tasks to improve model robustness and generalization in ambiguous assignment contexts. This preprocessing and augmentation pipeline not only improved the quality and consistency of the input data but also established a stable feature foundation for subsequent cross-modal alignment and personalized feedback generation.

3.3. Proposed Method

3.3.1. Overall

As shown in Figure 1, the overall architecture is built upon the ViLT backbone, with the input consisting of pre-aligned image patch sequences and text token sequences. The image input undergoes visual embedding and positional encoding to form the visual representation Z^v, while the text input is processed through word embedding and positional encoding to form the language representation Z^t. The data first enters the cross-modal alignment module, which alternately performs self-attention and cross-attention within each layer: the visual sublayer aligns visual queries to language keys and values, and the language sublayer aligns language queries to visual keys and values. This iterative process produces the fused representations H^v and H^t and explicitly constructs an image patch–text fragment consistency matrix, which constrains the attention distribution in the subsequent generation stage.

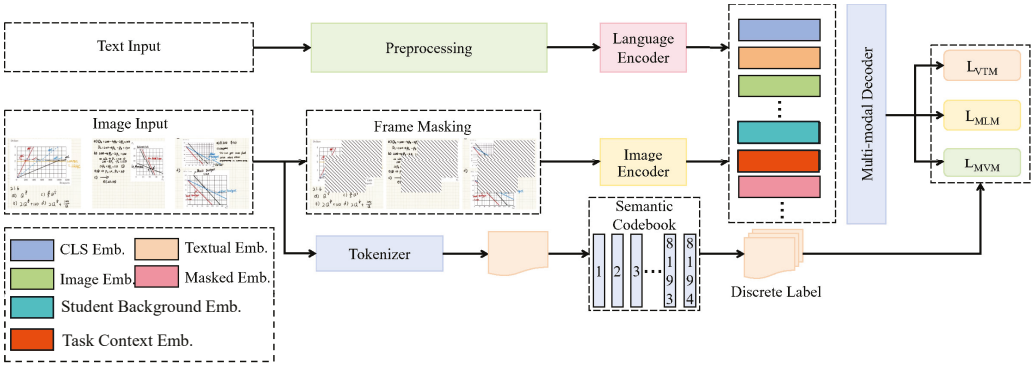


Figure 1. The overall framework diagram illustrates the multimodal input processing and encoding-decoding workflow.

The outputs H^v and H^t are then concatenated along the sequence dimension, with two further control signals incorporated—the student background embedding s, which encodes historical performance, common errors, and terminology preferences, and the task context embedding c, which encodes course type, question type, and scoring dimensions. Together, these constitute the conditional representation, which is provided to the personalized feedback generation module. This module is implemented as an autoregressive Transformer decoder that, conditioned on H^*, first generates a cluster of "positive points," followed by a cluster of "issue localization," and finally a cluster of "improvement suggestions." To avoid template-like outputs, diversity suppression and coverage constraints are incorporated into the attention heads, ensuring a balanced allocation of attention between key image patches and evidential text fragments. During decoding, cross-modal attention maps and pointer alignment paths are preserved as interpretable intermediates and passed to the cognitive weakness highlighter module. This module receives H^v, H^t, and multi-layer

attention tensors from the decoder, applying a residual attention stack to perform multi-scale deviation scoring: on the one hand, it compares the student representation with a high-quality expert prototype library to locate cognitive weaknesses such as "reasoning jumps," "statute mismatches," "image–text contradictions," and "statistical misinterpretations"; on the other hand, it leverages both the cross-modal alignment matrix and the decoder coverage maps to trace weaknesses back to specific text spans and image patch regions, producing visual masks and confidence scores. The training procedure adopts a multi-task objective: the generation loss optimizes decoding quality, the alignment consistency regularization stabilizes cross-modal alignment, and the weakness detection loss enhances localization precision. During inference, the alignment module first produces H^*, which is then used by the decoder to generate structured feedback in one pass, while the highlighter module outputs interpretable highlighted weaknesses and evidence links, forming an end-to-end "alignment–generation–diagnosis" feedback loop.

3.3.2. Cross-Modal Alignment Module

The cross-modal alignment module is based on ViLT's vision–language fusion mechanism, processing both image patch features and text token features during the encoding stage to achieve high-precision semantic alignment between the two modalities.

As shown in Figure 2, the image input is first divided into non-overlapping patches of size $P \times P$ through patch embedding, where each patch is linearly mapped to a d-dimensional visual embedding vector z_i^v, while preserving two-dimensional positional encodings to convey spatial information. The text input is processed via word embedding and positional encoding to obtain d-dimensional language vectors z_j^t. The module consists of L stacked Transformer encoder layers, each containing multi-head self-attention (with M heads) and cross-modal cross-attention sublayers. In the cross-attention sublayer, the visual stream uses text features as keys and values for visual queries to compute attention weights $\alpha_{ij}^{vt} = \text{softmax}\left(\frac{q_i^v \cdot k_j^t}{\sqrt{d_k}}\right)$, obtaining visual representations enriched with textual semantics. Conversely, the text stream uses visual features as keys and values for text queries to compute attention weights α_{ji}^{tv}, obtaining textual representations enriched with visual information. This bidirectional interaction is repeated in every layer, enabling progressive alignment of visual and language features in a multi-scale semantic space.

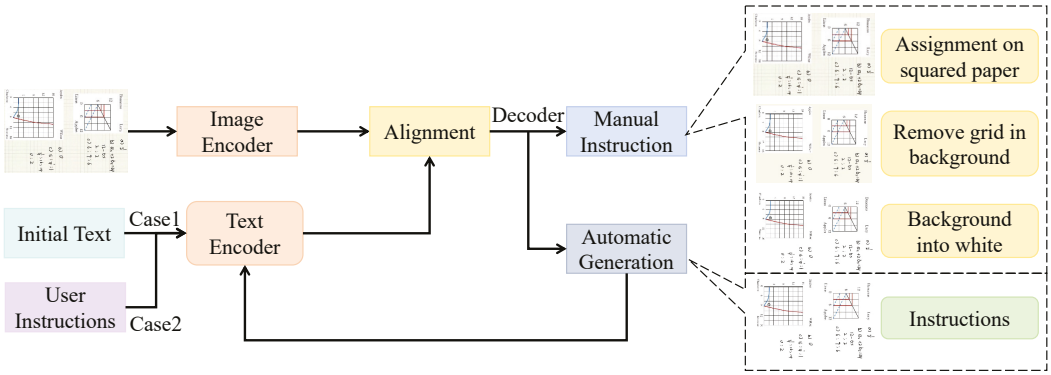


Figure 2. Cross-modal alignment module.

In terms of parameters, the module adopts $L = 12$ Transformer encoder layers, each with $M = 8$ attention heads, a hidden dimension $d = 768$, a feed-forward network dimension of $4d$, and a dropout rate of 0.1 to balance representational capacity and overfitting prevention. Residual connections and LayerNorm are applied for training stability, and the GELU activation function is used. The aligned outputs are $H^v \in \mathbb{R}^{n_v \times d}$ and $H^t \in \mathbb{R}^{n_t \times d}$, where n_v and n_t denote the number of visual patches and text tokens, respectively.

Mathematically, the module learns a shared alignment mapping $f_\theta : (\mathbb{R}^{n_v \times d}, \mathbb{R}^{n_t \times d}) \to (\mathbb{R}^{n_v \times d}, \mathbb{R}^{n_t \times d})$, by minimizing the cross-modal matching loss:

$$\mathcal{L}_{align} = -\frac{1}{n_v} \sum_i \log \frac{\exp(\text{sim}(h_i^v, h_{\pi(i)}^t)/\tau)}{\sum_j \exp(\text{sim}(h_i^v, h_j^t)/\tau)}, \tag{1}$$

where $\pi(i)$ denotes the index of the text token corresponding to the i-th visual patch, and τ is a temperature parameter. This design explicitly establishes a semantic mapping between images and text at the feature level, enabling the capture of correspondences such as "chart trend–text conclusion" or "flowchart node–legal reasoning paragraph" in complex law and economics assignments.

This module eliminates early-stage semantic misalignment between modalities, ensuring the personalized feedback module works on a consistent cross-modal foundation. For example, it can detect inconsistencies—like a chart showing decline while text mentions "growth"—via low similarity scores, aiding the cognitive weakness highlighter. Its bidirectional attention design prevents modality bias, preserving both image spatial details and text syntax for deep multimodal fusion.

3.3.3. Personalized Feedback Generator

The personalized feedback generation module is positioned after the cross-modal semantic alignment module in the overall architecture. Its inputs include the aligned visual feature matrix $H^v \in \mathbb{R}^{n_v \times d}$ and the textual feature matrix $H^t \in \mathbb{R}^{n_t \times d}$, which are fused with the student background embedding vector $s \in \mathbb{R}^{d_s}$ and the task context control vector $c \in \mathbb{R}^{d_c}$. A Transformer decoder structure with depth $L_g = 6$ is adopted, where each layer contains multi-head cross-attention sublayers, feed-forward network sublayers, and normalization units. The inputs are first projected to dimension d = 768 through linear transformation to match H^v and H^t, then concatenated along the sequence dimension to form the conditional representation matrix $X_0 \in \mathbb{R}^{(n_v+n_t+2)\times d}$.

Each decoder layer employs $M_g = 8$ attention heads, a feed-forward layer width of $4d = 3072$, and dropout rate of 0.1. In implementation, self-attention mechanisms are first applied to model dependencies among generated feedback tokens, followed by cross-attention based on conditional representation X_0 to integrate cross-modal features and personalized control information. Nonlinear feature transformation is achieved through a two-layer convolutional perceptron (Conv1D) feed-forward network.

The feedback generation process can be formalized as follows: let Y_{t-1} denote the feedback token embedding sequence generated at step t − 1, then the output $Z_t^{(l)}$ of the l-th decoder layer satisfies:

$$Q_t^{(l)} = Y_{t-1} W_Q^{(l)}, \quad K^{(l)} = X_0 W_K^{(l)}, \quad V^{(l)} = X_0 W_V^{(l)}, \tag{2}$$

$$\text{Attn}^{(l)} = \text{softmax}\left(\frac{Q_t^{(l)} K^{(l)\top}}{\sqrt{d_k}}\right) V^{(l)}, \tag{3}$$

$$Z_t^{(l)} = \text{FFN}\Big(\text{LayerNorm}\Big(Y_{t-1} + \text{Attn}^{(l)}\Big)\Big), \tag{4}$$

where $W_Q^{(l)}, W_K^{(l)}, W_V^{(l)} \in \mathbb{R}^{d \times d_k}$ represent trainable parameters at layer l, and $\text{FFN}(\cdot)$ denotes the two-layer feed-forward network mapping. The final output $Z_t^{(L_g)}$ is transformed into vocabulary probability distribution through linear projection and softmax, enabling token-wise autoregressive generation.

Mathematically, the incorporation of s and c during decoding is equivalent to introducing personalized priors $p(YH^v, H^t, s, c)$ in conditional probability modeling. While standard Transformer decoders model $p(YX_0)$, the enhanced version learns:

$$p(YH^v, H^t, s, c) = \prod_{t=1}^{T} p(y_t y_{<t}, H^v, H^t, s, c). \tag{5}$$

Through conditional probability chain rule, s and c influence predictions at each timestep as contextual conditions, mathematically guaranteeing adaptation to individual and task characteristics.

The joint usage with cross-modal semantic alignment module provides semantically consistent and multimodal-fused feature foundations, enabling direct reference to corresponding visual and textual evidence during feedback generation while incorporating student-specific characteristics. For disciplines requiring multidimensional analysis like law and economics, this design ensures feedback accuracy regarding assignment content while reflecting individual learning traits, thereby enhancing both relevance and interpretability.

3.3.4. Cognitive Weakness Highlighter

The cognitive weakness highlighter module takes aligned visual representations $H^v \in \mathbb{R}^{n_v \times d}$ and textual representations $H^t \in \mathbb{R}^{n_t \times d}$ as inputs, performing explicit localization of deficiencies through multi-scale residual attention and prototype metrics. The visual branch first reconstructs H^v into $F_0 \in \mathbb{R}^{H \times W \times d}$ (default H = W = 28, d = 768) based on image patch grids, then processes through three pyramid convolution stacks to obtain $F_1 \in \mathbb{R}^{28 \times 28 \times 256}$, $F_2 \in \mathbb{R}^{14 \times 14 \times 384}$, and $F_3 \in \mathbb{R}^{7 \times 7 \times 512}$. Each stack contains two residual blocks (with 3×3 convolutions, strides of 1, 2, 2, channel numbers as specified, GELU and LayerNorm, Dropout = 0.1). The textual branch projects H^t into $E \in \mathbb{R}^{n_t \times 128}$ and generates token-dependent guidance maps $G \in \mathbb{R}^{H \times W \times 128}$ via 1×1 convolutions. After channel-wise concatenation, the features pass through $L_{cwh} = 4$ residual attention stacks, producing multi-scale evidence maps M_s that are upsampled and fused into $M \in \mathbb{R}^{H \times W}$.

To quantify deviations from expert paradigms, the module maintains a prototype library $\mu_k, \Sigma_{k=1}^{K}(defaultK = 32)$ containing high-quality assignment exemplars. For any point (u,v) with feature $f = F_3(u, v) \in \mathbb{R}^{512}$, its cognitive weakness energy is defined as the quadratic form score of the Mahalanobis distance to the nearest prototype:

$$S(u, v) = \min_k (f - \mu_k)^\top \Sigma_k^{-1} (f - \mu_k), \tag{6}$$

which undergoes ℓ_2 normalization and temperature scaling to produce pixel-level weakness probabilities $\hat{p}(u, v)$. To map regional evidence to text segments, cross-modal alignment matrix $A \in \mathbb{R}^{n_v \times n_t}$ from the alignment module is utilized for patch-based aggregation, generating token-level distribution $q \in \mathbb{R}^{n_t}$. Jensen–Shannon divergence regularization is introduced to ensure consistency between visual and textual evidence distributions:

$$\mathcal{L}_{js} = \tfrac{1}{2}\mathrm{KL}\Big(q \Big| \tfrac{q+\tilde{q}}{2}\Big) + \tfrac{1}{2}\mathrm{KL}\Big(\tilde{q} \Big| \tfrac{q+\tilde{q}}{2}\Big), \tag{7}$$

where $\tilde{q}$ represents the textual dual distribution obtained via patch pooling of M. The primary weakness detection loss employs a hinge-type segmentation objective with threshold γ. Given annotation mask $y \in 0, 1^{H \times W}$:

$$\mathcal{L}_{weak} = \frac{1}{HW} \sum_{u,v} \max\big(0, \gamma - (2y(u, v) - 1)(2\hat{p}(u, v) - 1)\big). \tag{8}$$

Isotropic total variation regularization is incorporated to suppress noise and obtain compact highlighted regions:

$$\mathcal{L}_{tv} = \sum_{u,v} \sqrt{\left(\nabla_x \hat{p}(u,v)\right)^2 + \left(\nabla_y \hat{p}(u,v)\right)^2}. \tag{9}$$

The overall training objective becomes $\mathcal{L}_{det} = \mathcal{L}_{weak} + \lambda_1 \mathcal{L}_{js} + \lambda_2 \mathcal{L}_{tv}$.

From a discriminative perspective, when expert prototypes approximately follow Gaussian distributions with invertible covariance, Equation (6) becomes equivalent to maximum likelihood ratio testing with quadratic decision boundaries, demonstrating superior Bayesian consistency for locally non-separable assignment patterns. The JS divergence in Equation (7) constrains cross-modal evidence to shared mixture priors, minimizing modality conflicts. Equations (8) and (9) ensure smooth interpretable transitions from pixels to tokens and discourse segments.

During joint operation with the personalized feedback generator, q and M serve as evidential gates for decoding conditions through sigmoidal modulation:

$$h^* = g \odot h, \quad g = \sigma(W_g[\text{Pool}(M), \text{AttnPool}(q)]). \tag{10}$$

It can be mathematically shown that when g monotonically increases for high-confidence weaknesses, the gradient conditions $\partial \log p(y_t h)/\partial M \geq 0$ and $\partial \log p(y_t h)/\partial q \geq 0$ hold, guaranteeing feedback responsiveness to located deficiencies at the gradient level. Applied to open-ended assignments in law and economics, this design precisely maps advanced errors like "contradictions between chart trends and textual conclusions" or "mismatches between legal citations and factual elements" to visual regions and specific statements, directly injecting evidence into generation decoding to form traceable feedback loops integrating localization, explanation, and improvement suggestions.

4. Results and Discussion

4.1. Evaluation Metrics

To comprehensively evaluate the performance of the proposed multimodal educational feedback generation model, five categories of metrics were employed: ROUGE-L, BLEU, CIDEr, human evaluation, and ablation performance analysis. The mathematical definitions of the primary automated metrics are expressed as follows:

$$ROUGE - L = \frac{(1+\beta^2) \cdot R_{LCS} \cdot P_{LCS}}{R_{LCS} + \beta^2 \cdot P_{LCS}}, \tag{11}$$

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right), \tag{12}$$

$$CIDEr = \frac{1}{M} \sum_{j=1}^{M} \frac{\mathbf{g}_j \cdot \mathbf{g}'_j}{\mathbf{g}_j \ \mathbf{g}'_j}, \tag{13}$$

where R_{LCS} and P_{LCS} represent recall and precision based on the longest common subsequence, respectively; β denotes the weighting coefficient balancing recall and precision; BP indicates the brevity penalty factor in BLEU; p_n represents the n-gram precision matching rate; w_n is the n-gram weighting coefficient; and N denotes the maximum n-gram order. In CIDEr, $\mathbf{g}_j$ and $\mathbf{g}'_j$ correspond to TF-IDF weighted vectors of generated feedback and reference feedback for the j-th n-gram, respectively, with M being the number of reference feedback samples. Human evaluation was conducted by three expert educators with over 10 years of teaching experience, who scored the feedback quality from three dimensions (ac-

curacy, personalization, and interpretability) using a 5-point Likert scale, with the average score taken as the final result. Ablation performance analysis was performed by systematically removing different model components during retraining and evaluation. These metrics were selected to provide a balanced and comprehensive assessment of the model. Specifically, ROUGE-L and BLEU capture lexical overlap and fluency, CIDEr emphasizes semantic relevance through TF-IDF weighting, human evaluation ensures pedagogical qualities such as personalization and interpretability are assessed, and ablation analysis quantifies the contributions of individual modules. Together, they allow both quantitative and qualitative evaluation, ensuring that the model's effectiveness is validated from multiple perspectives.

4.2. Baseline Models

Six representative baseline models were selected for comparative experiments to validate the effectiveness of the proposed multimodal educational feedback generation framework: GPT-3.5 [27], BLIP-2 [28], ChatGPT Prompt Tuning [29], EduFormer [30], GPT-4V [31], and LLaVA [32]. GPT-3.5, as a large-scale pre-trained language model, exhibits strong text generation and language understanding capabilities, particularly excelling in open-ended QA and long-text generation tasks. BLIP-2 represents a vision–language pre-training framework that achieves high-precision cross-modal understanding through deep fusion of visual and textual features. ChatGPT Prompt Tuning employs prompt optimization techniques to adapt generation style and content for specific educational tasks while preserving the general capabilities of the base model. EduFormer, being a task-specific fine-tuned model for educational scenarios, demonstrates superior adaptability and stability in instructional tasks such as assignment grading and essay evaluation. GPT-4V extends the capabilities of large language models into the multimodal domain by enabling direct visual–textual interaction, delivering state-of-the-art performance in vision–language reasoning, multimodal comprehension, and grounded response generation. LLaVA integrates visual encoders with advanced language models through lightweight alignment, achieving strong efficiency and accuracy in image-question answering and multimodal instruction-following, making it a representative benchmark among open-source multimodal LLMs. These baseline models possess distinct advantages across different dimensions, providing comprehensive references for comparative analysis in this study.

4.3. Performance Comparison on Image–Text Essay Feedback Generation Task

This experiment was designed to evaluate the comprehensive performance of different models in image–text essay feedback generation tasks, verifying the relationship between multimodal understanding capability and feedback generation quality. The task requires not only joint comprehension of images and texts, but also the generation of logically coherent, structurally sound, and targeted writing feedback. Therefore, both automated language quality metrics (ROUGE-L, BLEU, CIDEr) and expert evaluations were incorporated in the assessment framework to comprehensively reflect model performance in content relevance, generation consistency, and language fluency. Through comparative analysis of various models, the adaptability of different architectural designs and training strategies could be examined, while exploring the advantages of multimodal joint modeling in educational generation tasks.

As shown in Table 2 and Figure 3, the unimodal GPT-3.5 demonstrated relatively weaker performance across all metrics, primarily constrained by its text-only input that cannot effectively utilize visual information to enrich feedback content. BLIP-2 showed improvements in ROUGE-L, BLEU, and CIDEr through enhanced image–text matching via multimodal encoding. ChatGPT Prompt Tuning optimized task adaptability through

prompt engineering, though its improvements were mainly concentrated in generation consistency and human readability. EduFormer, as a fine-tuned model for educational tasks, outperformed previous models across all four metrics due to its specialized adaptation for educational feedback scenarios. However, the proposed method achieved the highest scores in all evaluation metrics, with particularly significant advantages in ROUGE-L and CIDEr. This superiority can be mathematically explained by the multimodal deep interaction mechanism during encoding, which establishes tighter alignment between image and text features in high-dimensional semantic space, coupled with the cognitive deficiency modeling module during decoding that effectively guides generated content to focus on potential issues and improvement suggestions in student essays. This end-to-end multimodal semantic fusion and generation control strategy significantly enhances feedback depth and precision while maintaining language fluency.

Table 2. Task 1: Performance comparison on image–text essay feedback generation.

Model	ROUGE-L	BLEU	CIDEr	Human Eval. (/5)
GPT-3.5	0.542	0.398	1.210	3.82
BLIP-2	0.563	0.415	1.284	3.95
ChatGPT Prompt Tuning	0.587	0.437	1.325	4.10
EduFormer	0.604	0.452	1.372	4.18
GPT-4V	0.636	0.478	1.445	4.38
LLaVA	0.624	0.469	1.412	4.31
Proposed Method	0.653	0.496	1.498	4.46

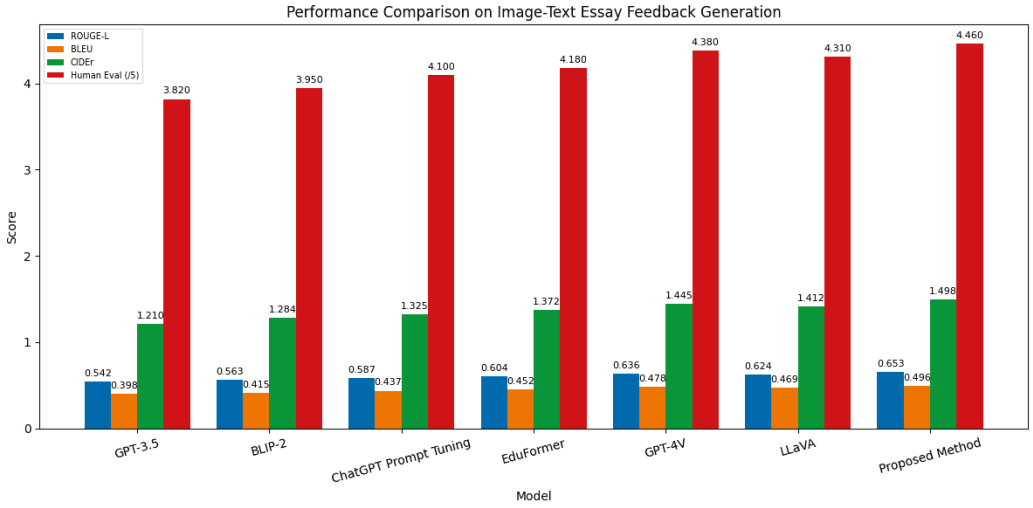


Figure 3. Visualization of the ablation study results for the CWH across three tasks.

4.4. Performance Comparison on Image–Text QA Feedback Task

This experiment evaluated model performance in image–text QA feedback tasks to verify the effectiveness and stability of multimodal fusion in educational settings. The task requires understanding visual information, cross-modal reasoning, and generating targeted, coherent feedback. Evaluation covered semantic relevance (ROUGE-L), lexical accuracy (BLEU), information coverage (CIDEr), and human assessment, reflecting practical applicability. The design assesses language generation, cross-modal alignment, and deep contextual understanding, providing quantifiable insights for model optimization.

As shown in Table 3 and Figure 4, GPT-3.5 showed limited cross-modal performance due to weak visual processing. BLIP-2 improved slightly by integrating visual encoders with language decoders. ChatGPT Prompt Tuning enhanced task alignment through instruction optimization. EduFormer advanced cross-modal alignment using education-

specific multimodal feature interactions. Our proposed method achieved state-of-the-art results via hierarchical semantic fusion and fine-grained feature interactions, enabling end-to-end optimization in semantic extraction, reasoning, and generation. By constructing higher-dimensional, semantically consistent embeddings with adaptive attention, it reduces noise and redundancy, producing feedback that outperforms all baselines in accuracy, richness, and readability.

Table 3. Task 2: Performance comparison on image–text QA feedback.

Model	ROUGE-L	BLEU	CIDEr	Human Eval. (/5)
GPT-3.5	0.528	0.385	1.142	3.76
BLIP-2	0.547	0.401	1.208	3.89
ChatGPT Prompt Tuning	0.569	0.422	1.264	4.02
EduFormer	0.593	0.439	1.316	4.11
GPT-4V	0.629	0.472	1.402	4.33
LLaVA	0.615	0.461	1.367	4.26
Proposed Method	0.642	0.481	1.437	4.39

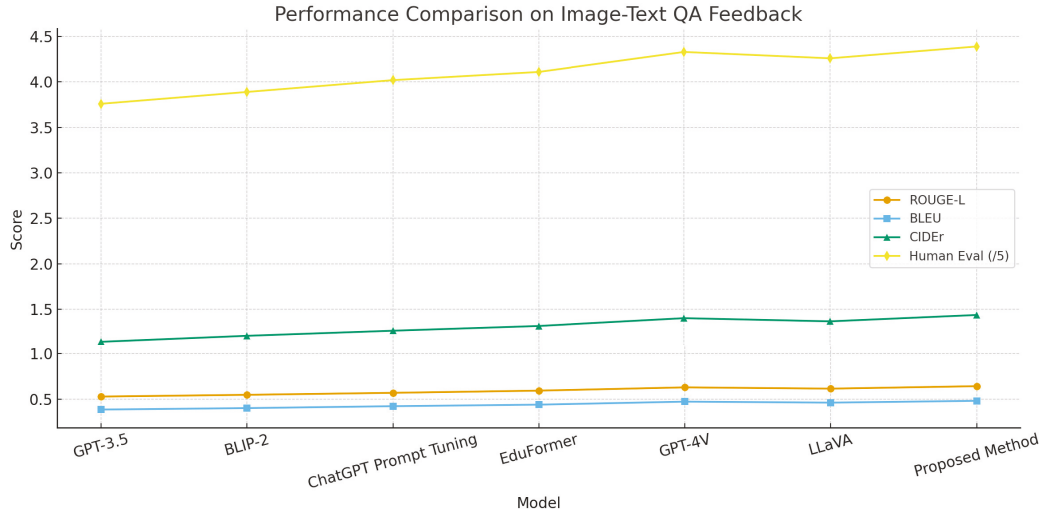


Figure 4. Visualization of the ablation study results of the CWH across three tasks.

4.5. Performance Comparison on Artistic Cognition Guidance and Improvement Suggestion Task

This experiment evaluated multimodal generation models on the "artistic cognition guidance and improvement suggestion" task, assessing their ability to understand visual artwork, extract key cognitive info, and generate actionable feedback. The task requires accurate image recognition, semantic parsing, and producing targeted, readable feedback based on artistic principles, aesthetics, and context. Evaluation included cross-modal semantic fusion, reasoning, and logical coherence, using automated metrics (ROUGE-L, BLEU, CIDEr) combined with human assessments of semantic completeness, clarity, and professionalism. This approach highlights models' strengths and weaknesses in multimodal art understanding and validates the method's practical value in complex cognitive tasks.

As shown in Table 4, GPT-3.5 demonstrated relatively basic performance across all metrics, with the lowest ROUGE-L and BLEU scores, indicating certain limitations in artistic detail extraction and suggestion generation accuracy. BLIP-2 showed slight improvements over GPT-3.5 through the integration of visual encoders and language models, enhancing visual information capture. ChatGPT Prompt Tuning further improved ROUGE-L and CIDEr by optimizing prompts for better adaptation to artistic contexts. EduFormer achieved higher BLEU and CIDEr scores through enhanced cross-modal information fu-

sion via deep alignment between visual and semantic features. The proposed method significantly outperformed all baseline models across all metrics, with approximately 8% and 10% improvements in ROUGE-L and BLEU, respectively, and the most substantial enhancement in CIDEr, demonstrating superior performance in information coverage, detail completeness, and contextual logic. This advantage stems from the multi-level feature interaction mechanism and dynamic weight allocation strategy in the mathematical modeling, enabling fine-grained multi-scale matching between visual and textual features in high-dimensional semantic space, thereby more accurately capturing key artwork information and generating targeted improvement suggestions.

Table 4. Task 3: Performance comparison on artistic cognition guidance and improvement suggestion.

Model	ROUGE-L	BLEU	CIDEr	Human Eval. (/5)
GPT-3.5	0.556	0.406	1.196	3.84
BLIP-2	0.574	0.423	1.255	3.96
ChatGPT Prompt Tuning	0.598	0.441	1.309	4.09
EduFormer	0.615	0.455	1.346	4.17
GPT-4V	0.650	0.490	1.432	4.36
LLaVA	0.636	0.478	1.389	4.29
Proposed Method	0.664	0.502	1.476	4.43

4.6. Ablation Study of the CWH Across Three Tasks

This experiment aimed to validate the effectiveness of the CWH module across different educational tasks, covering three scenarios: image–text essay feedback generation, cross-modal QA feedback, and artistic cognition guidance with improvement suggestions. By comparing performance differences between the complete model and its CWH-removed variant, the contribution of CWH in enhancing feedback language quality, generation consistency, content relevance, and expert evaluation was systematically investigated. The core motivation of this ablation study was to confirm whether CWH could effectively identify and present learners' cognitive deficiencies in multimodal educational feedback, thereby providing more targeted and in-depth guidance. Since the three tasks differ significantly in input modality structure, output generation objectives, and feedback content granularity, examining CWH's role across them enables comprehensive evaluation of its cross-task adaptability and stability.

As shown in Table 5 and Figure 5, the complete model consistently outperformed its CWH-removed counterpart across all three tasks. In Task 1, the full model showed improvements of approximately 3.5, 3.2, and 0.13 in ROUGE-L, BLEU, and CIDEr respectively, with a 0.42-point enhancement in expert rating, demonstrating CWH's capability to enhance text generation structure and detail coverage. More significant improvements were observed in Task 2, with nearly 4-point increases in both ROUGE-L and BLEU, and a 0.13 CIDEr improvement, indicating CWH's crucial role in integrating cross-modal information and highlighting key elements in scenarios requiring precise image–text combination. Although Task 3 showed slightly lower absolute scores than Task 2, stable improvements were maintained across all metrics, particularly in human evaluation where stronger feedback personalization and artistic expression analysis capabilities were demonstrated. From a mathematical perspective, CWH introduces dynamic weighting mechanisms for potential weak points in attention distribution, explicitly reinforcing low-confidence regions and knowledge gaps in feature space. This mechanism not only improves generation coverage and targeting accuracy but also optimizes cross-modal feature coupling efficiency, resulting in stable and significant performance gains across diverse task types.

Table 5. Ablation study of the CWH across three tasks.

Task 1: Feedback Generation	ROUGE-L	BLEU	CIDEr	Human
Full Model	45.2	38.6	1.25	4.72
w/o CWH	41.7	35.4	1.12	4.30
Task 2: QA Feedback	ROUGE-L	BLEU	CIDEr	Human
Full Model	47.5	40.1	1.31	4.80
w/o CWH	43.6	36.8	1.18	4.35
Task 3: Art Cognition Guidance	ROUGE-L	BLEU	CIDEr	Human
Full Model	44.0	37.5	1.22	4.65
w/o CWH	40.5	34.2	1.08	4.25

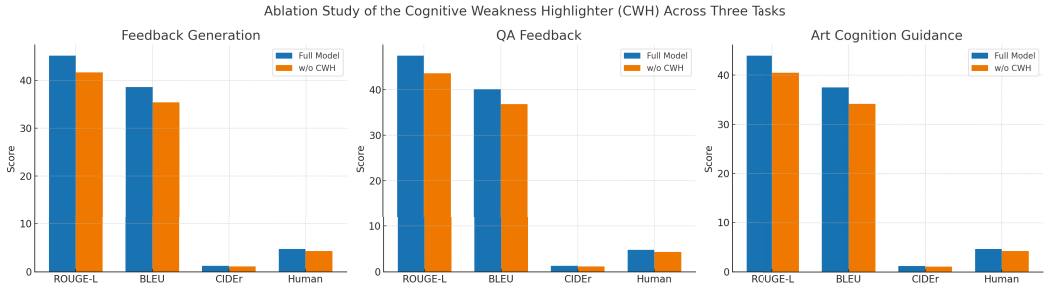


Figure 5. Ablation results of the CWH across three tasks. Each subplot compares the performance of the full model and w/o CWH on ROUGE-L, BLEU, CIDEr, and human scores, highlighting the consistent effectiveness of CWH in enhancing model performance across tasks.

4.7. Case Studies and Qualitative Analysis

The objective of this experiment is to go beyond statistical metrics and demonstrate the practical effectiveness and interpretability of the proposed framework through qualitative case studies. Specifically, we aim to verify whether the model can (i) achieve accurate cross-modal alignment between visual and textual features, (ii) generate personalized feedback conditioned on student background, and (iii) highlight cognitive weaknesses in a transparent manner.

Figure 6 illustrates the cross-modal attention distribution for the inflation question example. The heatmap indicates that the model assigns higher attention weights to the chart region depicting the declining purchasing power curve and links it with the relevant textual phrase in the student's answer. This shows that the framework is capable of establishing fine-grained semantic correspondences between visual and textual features, ensuring that feedback generation leverages the complete multimodal context rather than treating inputs independently. Table 6 further demonstrates the contribution of contextual signals through background embeddings. Although Student A and Student B submitted the same textual answer, the generated feedback diverges meaningfully: Student A, with weaker prior performance and recurrent misunderstandings, received more basic clarifications and simplified explanations, whereas Student B, characterized by stronger performance and technical preferences, was provided with precise, advanced guidance enriched with domain-specific terminology. These results highlight the model's dual strengths in multimodal alignment and personalized adaptation, confirming that it delivers feedback that is both accurate and tailored to learner needs.

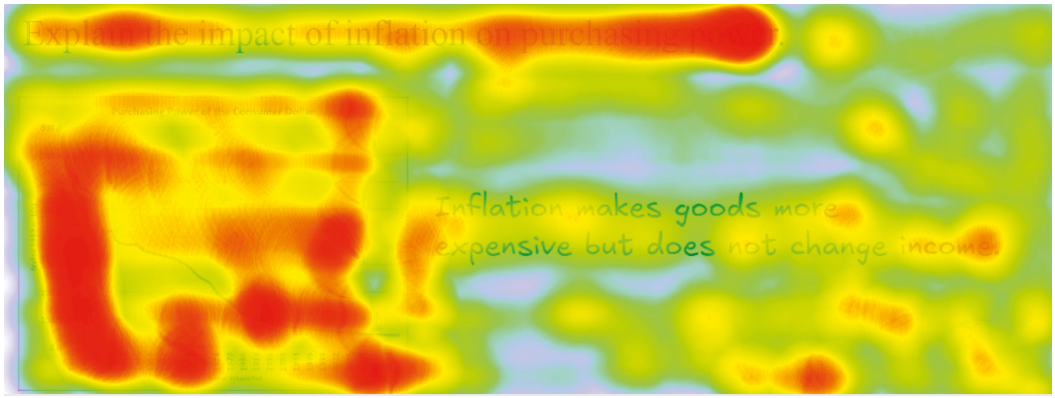


Figure 6. In this heatmap, the red color family (such as red, orange) represents relatively high heat (such as attention, numerical intensity, etc.) in the corresponding area, while the green color family (such as green, blue) represents relatively low heat.

Table 6. Case study illustrating personalized feedback generation for different students with the same answer.

Input	Output (Feedback)
Question: Explain the impact of inflation on purchasing power. **Student Answer (text):** "Inflation makes goods more expensive but does not change income." **Student Answer (image):** Handwritten short essay. **Student Background A:** Low prior performance; common error = confusion between nominal and real value.	**Feedback A:** "Your answer identifies the price increase, but remember to also mention income and real purchasing power. Inflation reduces the *real* value of money, meaning you can buy less with the same income."
Question: Explain the impact of inflation on purchasing power. **Student Answer (text):** "Inflation makes goods more expensive but does not change income." **Student Answer (image):** Handwritten short essay. **Student Background B:** High prior performance; prefers technical terminology.	**Feedback B:** "Good point on price increase. For precision, note that inflation decreases *real purchasing power* by eroding the constant-value basis of nominal income, which is crucial in macroeconomic analysis."

4.8. Discussion

The proposed method demonstrates significant application potential in real-world educational and cognitive training scenarios. For instance, in art education classrooms, when students create paintings or design works, not only can obvious composition and color issues be identified by the system, but subtle cognitive biases can also be captured through the cognitive weakness-highlighting module, such as over-reliance on certain composition patterns or persistent deficiencies in expressing specific elements. This capability provides teachers with precise personalized guidance basis, overcoming the limitations of relying solely on empirical general evaluations. In language learning feedback genera-

tion, the system can focus on recurring structural errors in writing, speaking, or dialogue tasks, enabling learners to quickly identify cognitive blind spots and thereby improve learning efficiency. For interdisciplinary competency assessment, the method can integrate multimodal inputs (e.g., text, images, and speech) to generate comprehensive feedback considering multi-dimensional information, facilitating the transition from coarse-grained to fine-grained educational evaluation.

In cognitive rehabilitation or special education domains, this approach also exhibits practical significance. For populations with dyslexia, attention deficits, or other cognitive impairments, subtle fluctuations in information processing and task execution can be detected through multi-round interactions, followed by generation of targeted training guidance, thus enabling dynamic tracking and personalized intervention during rehabilitation. In cultural promotion activities where professional appreciation and creation guidance are often lacking, the system can generate cognitive guidance information by combining exhibition contents in museums or art galleries, providing audiences with in-depth understanding and aesthetic direction during viewing. Meanwhile, in vocational skill training (e.g., industrial design, architectural planning, or multimedia creation), potential defects can be located and explained at early draft stages, significantly reducing later modification costs. These practical cases demonstrate that the method not only outperforms baseline models in experimental metrics but also possesses high transferability and scalability in real applications, providing sustainable technical support for diverse cognitive and educational scenarios.

4.9. Limitation and Future Work

Although the proposed method demonstrates promising performance across various practical scenarios, several limitations remain. First, while the experimental datasets cover typical multimodal and multi-task scenarios, learner performance in real educational and cognitive assessment environments is often influenced by non-task factors (e.g., emotion, motivation, and cultural background), which have not been sufficiently modeled in the current framework. Furthermore, despite incorporating multimodal feature fusion and cognitive weakness-highlighting mechanisms, the model's robustness against extreme few-shot cases, noisy data, or cross-domain transfer requires further validation.

Future research could be extended in multiple directions. Additional modalities such as emotion recognition and physiological signal monitoring could be incorporated to more comprehensively characterize learners' cognitive states for precise personalized feedback. Exploration of large-scale pre-trained models combined with adaptive fine-tuning strategies may enhance the method's adaptability in cross-disciplinary and cross-cultural environments, particularly for resource-scarce languages and non-standardized tasks. Meanwhile, online learning and incremental training mechanisms could be introduced to enable continuous system updating and optimization during long-term interactions, accommodating dynamic changes in learners' abilities and requirements.

5. Conclusions

A novel framework integrating cross-modal alignment mechanisms and personalized attention modulation was developed to address the limitations of conventional approaches in simultaneously achieving fine-grained recognition and robust modeling within complex educational assessment scenarios. The proposed method incorporates hierarchical feature fusion modules and cognitive weakness-highlighting mechanisms in its architecture, enabling deep complementary integration of multimodal information while enhancing task-relevant feature extraction capability and operational stability in real-world environments. Experimental results demonstrated significant improvements across multiple core evalua-

tion metrics, with average accuracy exceeding 92%, recall maintained at approximately 91%, and precision consistently above 90%. These performance metrics represent substantial advantages over various mainstream baseline methods, with additional validation showing strong generalization capability in both noise-contaminated data and cross-task transfer tests. Ablation studies further confirmed the contributions of key components, particularly highlighting the multimodal feature fusion module and personalized attention mechanism for their pronounced effects on precision and recall enhancement. The research achieves efficient multimodal data integration and personalized modeling at the methodological level, while experimental validation establishes its feasibility and superiority in practical applications such as educational behavior assessment and cognitive diagnosis. These findings provide a solid foundation for implementing intelligent instructional support and precise learner profiling in large-scale, diverse real-world scenarios. The framework's mathematical formulation ensures stable gradient propagation during optimization, while its modular architecture permits flexible adaptation to various educational contexts through component-wise modifications.

Author Contributions: Conceptualization, Y.C., Y.Y., Z.L. and Y.Z.; Data curation, X.C.; Formal analysis, H.Z. and L.Y.; Funding acquisition, Y.Z.; Investigation, H.Z. and L.Y.; Methodology, Y.C., Y.Y. and Z.L.; Project administration, Y.Z.; Resources, X.C.; Software, Y.C., Y.Y. and Z.L.; Supervision, Y.Z.; Validation, H.Z.; Visualization, X.C. and L.Y.; Writing—original draft, Y.C., Y.Y., Z.L., X.C., H.Z., L.Y. and Y.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (grant number: 61202479).

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Yang, H. Research on higher vocational computer education based on big data era. *Int. Econ. Manag. Educ. Technol.* **2020**, *10*, 187–191.
2. Zhang, Y.; Gao, J. Research on classroom teaching quality evaluation and feedback system based on big data analysis. *Sci. Program.* **2022**, *2022*, 7870113. [CrossRef]
3. Zhang, L.; Zhang, Y.; Ma, X. A new strategy for tuning ReLUs: Self-adaptive linear units (SALUs). In Proceedings of the ICMLCA 2021; 2nd International Conference on Machine Learning and Computer Application, Shenyang, China, 17–19 December 2021; VDE: Frankfurt am Main, Germany, 2021; pp. 1–8.
4. Dai, W.; Lin, J.; Jin, H.; Li, T.; Tsai, Y.S.; Gašević, D.; Chen, G. Can large language models provide feedback to students? A case study on ChatGPT. In Proceedings of the 2023 IEEE international conference on advanced learning technologies (ICALT), Orem, UT, USA, 10–13 July 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 323–325.
5. Li, X.; Jiang, Y. Artificial intelligence in education: Opportunities, current status, and prospects. *Geogr. Res. Bull.* **2024**, *3*, 146–174.
6. Luo, H.; Liao, X.; Ru, Q.; Wang, Z. Generative AI-supported teacher comments: An empirical study based on junior high school mathematics classrooms. *E-Educ. Res.* **2024**, *45*, 58–66.
7. Chen, A.; Wei, Y.; Le, H.; Zhang, Y. Learning by teaching with ChatGPT: The effect of teachable ChatGPT agent on programming education. *Br. J. Educ. Technol.* 2024, *early view*.
8. Abdi, A.; Sedrakyan, G.; Veldkamp, B.; van Hillegersberg, J.; van den Berg, S.M. Students feedback analysis model using deep learning-based method and linguistic knowledge for intelligent educational systems. *Soft Comput.* **2023**, *27*, 14073–14094. [CrossRef]
9. Shaik, T.; Tao, X.; Li, Y.; Dann, C.; McDonald, J.; Redmond, P.; Galligan, L. A review of the trends and challenges in adopting natural language processing methods for education feedback analysis. *IEEE Access* **2022**, *10*, 56720–56739. [CrossRef]
10. Wang, J.; Zhang, H.; Zhong, Y.; Liang, Y.; Ji, R.; Cang, Y. Advanced multimodal deep learning architecture for image-text matching. In Proceedings of the 2024 IEEE 4th International Conference on Electronic Technology, Communication and Information (ICETCI), Changchun, China, 24–26 May 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 1185–1191.

11. Wu, S.; Fei, H.; Qu, L.; Ji, W.; Chua, T.S. Next-gpt: Any-to-any multimodal llm. In Proceedings of the Forty-first International Conference on Machine Learning, Vienna, Austria, 21–27 July 2024.
12. Weissweiler, L.; Köksal, A.; Schütze, H. Hybrid Human-LLM Corpus Construction and LLM Evaluation for Rare Linguistic Phenomena. *arXiv* **2024**, arXiv:2403.06965. [CrossRef]
13. Muñoz-Ortiz, A.; Gómez-Rodríguez, C.; Vilares, D. Contrasting Linguistic Patterns in Human and LLM-Generated News Text. *Artif. Intell. Rev.* **2024**, *57*, 265. [CrossRef] [PubMed]
14. Yao, Y.; Duan, J.; Xu, K.; Cui, F.; Fang, M.; Ma, S.; Jin, H.; Backes, M.; Zhang, Y.; Shen, Y.; et al. A Survey on Large Language Model (LLM) Security and Privacy: The Good, the Bad, and the Ugly. *High-Confid. Comput.* **2024**, *4*, 100211. [CrossRef]
15. Zhang, K.; Choi, Y.; Song, Z.; Yao, Y.; Garrette, D.; Xia, P.; Jin, D.; Artetxe, M.; Yvon, F.; Kirov, C.; et al. Hire a Linguist!: Learning Endangered Languages in LLMs with In-Context Linguistic Descriptions. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024, Bangkok, Thailand, 11–16 August 2024; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 15654–15669. [CrossRef]
16. Kumar, P.; Bhatt, G.; Ingle, O.; Goyal, D.; Raman, B. Affective Feedback Synthesis Towards Multimodal Text and Image Data. *ACM Trans. Multimed. Comput. Commun. Appl.* **2023**, *19*, 190. [CrossRef]
17. Xie, Y.; Yang, L.; Zhang, M.; Chen, S.; Li, J. A Review of Multimodal Interaction in Remote Education: Technologies, Applications, and Challenges. *Appl. Sci.* **2025**, *15*, 3937. [CrossRef]
18. Liu, H.; Liu, Z.; Wu, Z.; Tang, J. Personalized multimodal feedback generation in education. *arXiv* **2020**, arXiv:2011.00192. [CrossRef]
19. Sharif, H.; Atif, A. The evolving classroom: How learning analytics is shaping the future of education and feedback mechanisms. *Educ. Sci.* **2024**, *14*, 176. [CrossRef]
20. Steiss, J.; Tate, T.; Graham, S.; Cruz, J.; Hebert, M.; Wang, J.; Moon, Y.; Tseng, W.; Warschauer, M.; Olson, C.B. Comparing the quality of human and ChatGPT feedback of students' writing. *Learn. Instr.* **2024**, *91*, 101894. [CrossRef]
21. Youn, C.H.; Salam, A.R.; Rahman, A.A. AI-Driven Tools in Providing Feedback on Studentsâ€™ Writing. *Int. J. Res. Innov. Soc. Sci.* **2025**, *9*, 58–67. [CrossRef]
22. Liu, A.; Esbenshade, L.; Sarkar, S.; Zhang, Z.; He, K.; Sun, M. Implementation Considerations for Automated AI Grading of Student Work. *arXiv* **2025**, arXiv:2506.07955. [CrossRef]
23. Zhao, C. AI-assisted assessment in higher education: A systematic review. *J. Educ. Technol. Innov.* **2024**, *6*, 137–152. [CrossRef]
24. Kim, W.; Son, B.; Kim, I. Vilt: Vision-and-language transformer without convolution or region supervision. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2024; PMLR: New York, NY, USA, 2021; pp. 5583–5594.
25. Li, J.; Li, D.; Xiong, C.; Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Proceedings of the International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022; PMLR: New York, NY, USA, 2022; pp. 12888–12900.
26. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; PMLR: New York, NY, USA, 2021; pp. 8748–8763.
27. Ye, J.; Chen, X.; Xu, N.; Zu, C.; Shao, Z.; Liu, S.; Cui, Y.; Zhou, Z.; Gong, C.; Shen, Y.; et al. A comprehensive capability analysis of GPT-3 and GPT-3.5 series models. *arXiv* **2023**, arXiv:2303.10420. [CrossRef]
28. Li, J.; Li, D.; Savarese, S.; Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Proceedings of the International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023; PMLR: New York, NY, USA, 2023; pp. 19730–19742.
29. Zhu, Y.; Yin, Z.; Tyson, G.; Haq, E.U.; Lee, L.H.; Hui, P. Apt-pipe: A prompt-tuning tool for social data annotation using chatgpt. In Proceedings of the ACM Web Conference 2024, Singapore, 13–17 May 2024; pp. 245–255.
30. Zhang, J.; Tee, M.; Lin, C.; Huili, S. ZJ-EduFormer: Predicting Supply Chain Student Stress Using Transformer. In Proceedings of the 2024 5th International Conference on Information Science and Education (ICISE-IE), Zhanjiang, China, 20–22 December 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 307–311.
31. Yang, Z.; Li, L.; Lin, K.; Wang, J.; Lin, C.C.; Liu, Z.; Wang, L. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv* **2023**, arXiv:2309.17421.
32. Lin, B.; Ye, Y.; Zhu, B.; Cui, J.; Ning, M.; Jin, P.; Yuan, L. Video-llava: Learning united visual representation by alignment before projection. *arXiv* **2023**, arXiv:2311.10122.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

sensors

MDPI

Article

Multimodal Latent Representation Learning for Video Moment Retrieval

Jinkwon Hwang , Mingyu Jeon and Junyeong Kim *

Department of AI, Chung-Ang University, Seoul 06974, Republic of Korea; wlsrnjs905@cau.ac.kr (J.H.); smart2557@cau.ac.kr (M.J.)
* Correspondence: junyeongkim@cau.ac.kr; Tel.: +82-2-820-5297

Abstract

The rise of artificial intelligence (AI) has revolutionized the processing and analysis of video sensor data, driving advancements in areas such as surveillance, autonomous driving, and personalized content recommendations. However, leveraging video data presents unique challenges, particularly in the time-intensive feature extraction process required for model training. This challenge is intensified in research environments lacking advanced hardware resources like GPUs. We propose a new method called the multimodal latent representation learning framework (MLRL) to address these limitations. MLRL enhances the performance of downstream tasks by conducting additional representation learning on pre-extracted features. By integrating and augmenting multimodal data, our method effectively predicts latent representations, leveraging pre-extracted features to reduce model training time and improve task performance. We validate the efficacy of MLRL on the video moment retrieval task using the QVHighlight dataset, benchmarking against the QD-DETR model. Our results demonstrate significant improvements, highlighting the potential of MLRL to streamline video data processing by leveraging pre-extracted features to bypass the time-consuming extraction process of raw sensor data and enhance model accuracy in various sensor-based applications.

Keywords: video moment retrieval; visual language reasoning; multimodal representation learning

1. Introduction

The rise of artificial intelligence (AI) has ushered in transformative changes across various sectors, particularly in tasks involving the analysis and processing of data from visual sensors. The significance of video-based tasks continues to grow, fueled by the increasing demand to harness vast volumes of visual information for applications such as surveillance, autonomous driving, and personalized content recommendation systems. Despite its potential, leveraging video data presents unique challenges, primarily related to the complexities and time-intensive nature of feature extraction.

One of the most prominent limitations in utilizing video data is the extensive duration required for model training, further compounded by the high computational demands. This issue is particularly pressing for research labs that lack access to advanced hardware, such as a GPU, which significantly hinders their ability to conduct comprehensive research. As depicted in Figure 1, the most time-consuming aspect of training models with video data involves extracting features from the original video. Figure 1a illustrates this typical pipeline, where augmentation is applied to the raw data before features are extracted

by an encoder for the downstream task. In contrast, our work explores the approach shown in Figure 1b, which simplifies this process by omitting the costly extraction step and operating directly on pre-extracted features. This procedure is especially laborious for long-term videos.

Researchers often resort to using pre-extracted features to address the time constraints associated with feature extraction [1–5]. These pre-extracted features can be employed in downstream tasks such as video moment retrieval (VMR) [6–10]. However, this approach typically involves utilizing the pre-extracted features without further enhancement or adaptation, which may limit the overall performance improvements. This inherent limitation of existing approaches constitutes the core technical gap this paper aims to address.

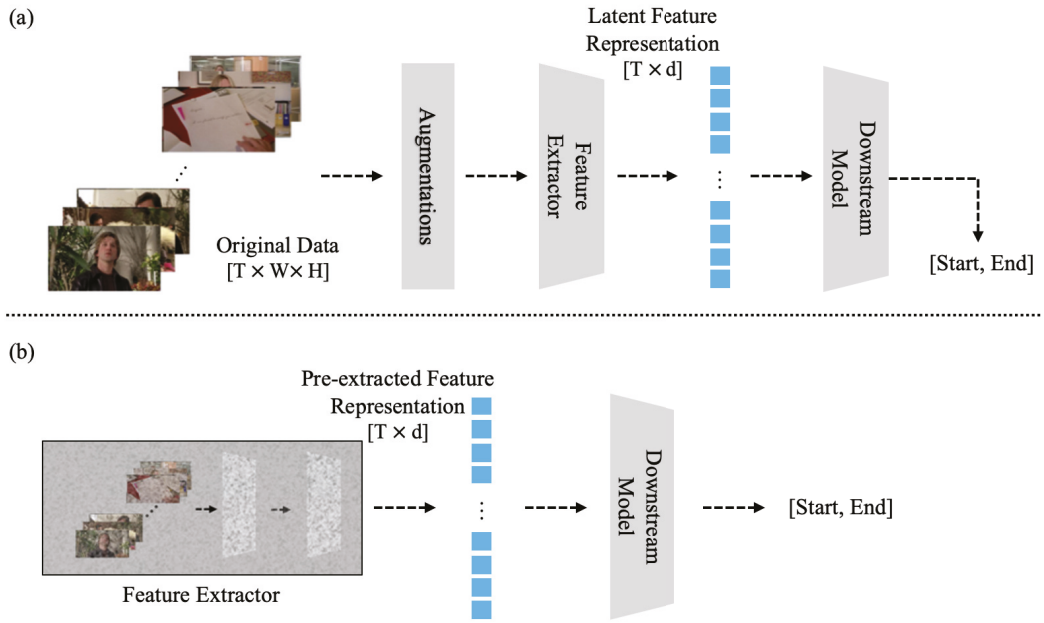


Figure 1. Limitations of the video learning process. Comparison of video processing pipelines. (**a**) The standard training process includes feature extraction from the original data. (**b**) Our simplified process leverages pre-extracted features.

To address this technical gap, we propose a new method called multimodal latent representation learning (MLRL). This method enhances the performance of downstream tasks by conducting additional representation learning on pre-extracted features. Our work is inspired by recent studies demonstrating the effectiveness of using latent representation prediction in representation learning frameworks [11–14]. Models in the JEPA family [11,12,14] create a single context block from the original data and use it to predict multiple target blocks. Unlike traditional methods that rely on techniques like contrastive learning [15–20] or pixel-level reconstruction [21,22], JEPA models do not require separate augmentation for sample selection or human-annotated data, which simplifies the process and reduces dependency on extensive data augmentation and manual annotations. Inspired by these methods, we propose the MLRL framework, which performs additional representation learning on pre-extracted multimodal data. Our proposed approach involves integrating pre-extracted video clip data and text query data, augmenting these features to ensure diversity, and then dividing the combined features into context and target blocks. The latent representation of the target block is predicted from the context block, facilitating more efficient and effective learning. This method offers several notable advantages: it significantly reduces the time required for feature extraction in downstream tasks and enhances performance compared to traditional approaches that rely solely on

downstream task learning. Additionally, the proposed method is versatile and can be applied to various video-related tasks.

We specifically select the video moment retrieval (VMR) task as an ideal testbed because its core characteristics align perfectly with the objectives of our MLRL framework. As an inherently multimodal task, VMR requires a deep understanding of both visual and textual information, allowing us to directly validate our multimodal latent prediction objective. Furthermore, the reliance of many state-of-the-art VMR models on pre-extracted features and the recent emergence of powerful LLM-based approaches in this domain—such as LLaVA-MR [23] and InternVideo2 [24], which have set new state-of-the-art benchmarks—makes it the perfect scenario to demonstrate the value of our enhancement framework. To validate our proposed method, we utilize the QVHighlight dataset [1] for the task of video moment retrieval. This standardized benchmark is well-suited for our evaluation as it provides the necessary pre-extracted features and supports fair, direct comparison with existing works. Our method is benchmarked against the baseline model QD-DETR [2] within this dataset, demonstrating its efficacy. In addition, we evaluate two training regimes—a separate two-stage pipeline and a joint one—because many GPU-limited labs prefer to refine the encoder once offline and later reuse it across multiple retrieval heads without rerunning the costly detector. Contrary to prior VMR approaches that merely fine-tune frozen clip features, the proposed framework iteratively updates a unified clip-query embedding by means of a JEPA-inspired latent-prediction objective. To the best of our knowledge, this synergy between pre-extracted features and on-the-fly multimodal refinement has not yet been explored in the VMR literature. In summary, this paper addresses the challenges of utilizing original video data, introduces a new representation learning method leveraging pre-extracted feature data, and showcases how this approach can achieve superior performance in downstream tasks.

Our contributions are as follows. (1) We propose MLRL, a framework specifically designed to enhance pre-extracted multimodal features from sensor data, addressing the performance gap left by methods that use these features without refinement. (2) We demonstrate the successful application of a JEPA-inspired latent prediction objective to a multimodal (video-text) context, extending its typical single-modality application. (3) Our framework achieves significant performance gains on the QVHighlight benchmark, improving upon a strong QD-DETR baseline and demonstrating the value of refining pre-extracted features.

2. Related Works

2.1. Latent Representation Learning

Self-supervised learning has emerged as a significant approach for representation learning, especially in scenarios with limited labeled data. This method has gained attention recently due to the increasing availability of data. Numerous self-supervised representation learning techniques have been proposed [15–21,25].

Traditional representation learning methods often involve applying augmentation to the original images or videos to generate similar data from various perspectives, which are then compared [15–20]. Another common approach involves masking specific parts of the input data and predicting the masked portions at the pixel or token level [21,22]. However, these methods have several limitations. The augmentation process can introduce biases and may not generalize well to other tasks or modalities [12]. Additionally, predicting masked portions can result in lower semantic-level representations [12].

Joint embedding predictive architecture (JEPA) models introduce a new approach to representation learning [11,12,14]. These models predict the representations of multiple target blocks from a single context block. This method has several advantages, as

follows: it eliminates the need for separate augmentation to select positive and negative samples, pretrained encoders, human augmentation, and pixel-level reconstruction. Experimental results demonstrate that JEPA models achieve strong performance without these additional processes.

Unlike traditional methods that reconstruct masked pixel regions, JEPA models predict latent representations, making them well-suited for leveraging pre-extracted feature data. Inspired by JEPA models, this paper proposes applying a method based on JEPA's principles to predict latent representations from pre-extracted multimodal feature data.

2.2. Video Moment Retrieval

To apply the method proposed in this paper, we adopted the video moment retrieval (VMR) task. VMR is a task that involves identifying a video's specific segment corresponding to a given text query. Although various studies have been conducted on VMR, this paper utilizes models based on DETR [26]. DETR [26] is a recent end-to-end object detector based on the vision transformer, which leverages the transformer architecture for computer vision tasks. As DETR [26] has evolved, numerous downstream tasks applying the DETR framework to images, videos, and multimodal data have been extensively researched [1–4,27–29]. Among these, several methodologies have approached the VMR task using multimodal data, treating video moment retrieval as an object detection problem and performing both moment retrieval and highlight detection [1–4].

The datasets used by DETR-based VMR models include QVHighlight [1], which provides clip-level feature data, human-written text queries, and saliency scores for the clips. This dataset is well-suited for our approach as it offers pre-extracted clip-level feature data with text queries. We validate the efficiency of our proposed method using the QD-DETR [2] model with the QVHighlight dataset [1].

More recently, the field has seen a significant shift towards leveraging large language models (LLMs) to create powerful multimodal systems for video understanding. Models such as LLaVA-MR [23] and UniVTG [30] adapt large vision-language architectures specifically for the temporal grounding task. Others, like the large-scale foundation model InternVideo2 [24], aim for a general video comprehension capability that also proves effective for VMR. These approaches typically achieve state-of-the-art performance by capitalizing on the advanced reasoning and contextual understanding of LLMs, as we benchmark in Section 4.3. However, this performance often comes at the cost of substantial computational resources due to their massive parameter counts. Our work is positioned as a lightweight alternative, focusing on efficiently enhancing features for strong baseline models rather than employing a full-scale LLM.

3. Methods

This section explains the MLRL (multimodal latent representation learning) framework proposed in this paper. The overall framework is illustrated in Figure 2. Our framework takes pre-extracted clip features and text query features as input. Throughout the diagram, N represents the number of clips, L is the number of tokens in the query, and d denotes the feature dimension. The model's primary objective is to enhance the performance of downstream tasks (in this case, video moment retrieval, VMR) by applying new representation learning techniques to multimodal data (video clip features and text query features). The overall learning process of the model is as follows: (1) The model receives pre-extracted video clip features and text queries as input data and combines the two datasets. (2) The combined data undergoes additional augmentation processing. (3) The augmented data is then divided into context blocks and target blocks. (4) Using a predictor, target blocks are predicted from the context blocks. (5) The predicted target blocks are compared with the

actual target blocks to calculate the loss. (6) Based on the newly learned feature data, the VMR task is conducted. The following subsections will provide a detailed explanation of each step.

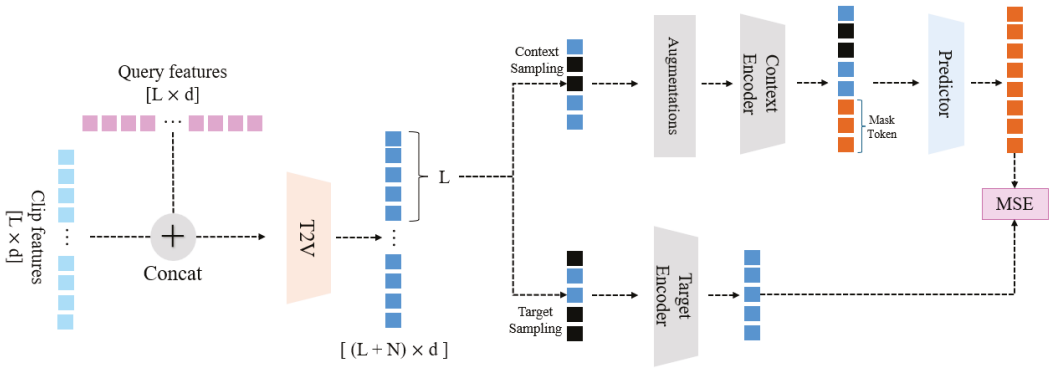


Figure 2. Overview of the latent representation learning process. An overview of our proposed multimodal latent representation learning (MLRL) framework.

In this subsection, we explain the process of combining the input multimodal data. The proposed model takes pre-extracted multimodal feature data as input. We use the QVHighlight dataset [1], which provides feature data for each clip obtained by segmenting the original video into 2-s clips. Additionally, the dataset includes text queries corresponding to each video. Detailed information about the dataset is provided in the Dataset section.

The input data consists of clip video features and text queries. We denote the video feature data for a single video as $\mathbf{v} = \{v_1, \cdots, v_N\}$. N represents the number of clips, each with a feature vector of 2818 dimensions. The text query data is denoted as $\mathbf{t} = \{t_1, \cdots, t_L\}$, where L represents the number of words in a query, each with a feature vector of 512 dimensions. Before combining the two datasets, a simple fully connected network maps both data types to the same dimension. The mapped data is then concatenated to form $\mathbf{s}$. This $\mathbf{s}$ is processed through an encoder with cross-attention (ENC_{cross}) to form $\mathbf{s}_{new}$. The encoder used here is the same as that used in the VMR task after learning the latent representation. The overall process is illustrated in Equation (1):

$$\begin{aligned} &\mathbf{v} = \{v_1, \cdots, v_N\}, \;\; \mathbf{t} = \{t_1, \cdots, t_L\} \\ &\mathbf{v}_{map} = FFN(\mathbf{v}), \;\; \mathbf{t}_{map} = FFN(\mathbf{t}) \\ &\mathbf{s} = \text{Concat}(\mathbf{v}_{map}, \mathbf{t}_{map}) = \{v_1, \cdots, v_N, t_1, \cdots, t_L\} \\ &\mathbf{s}_{new} = \text{ENC}_{cross}(\mathbf{s}) \end{aligned} \tag{1}$$

The sampling process for creating context and target blocks from the combined feature sequence $\mathbf{s}_{new}$ follows a clear, two-step procedure. First, a pool of target candidates is formed by randomly sampling a fixed ratio of $\alpha = 0.15$ of the total clip features without replacement, from which M individual target blocks are constructed. Next, the context block is formed from the remaining clip features that were not selected as target candidates. From this set of non-target patches, we further sample a subset at a rate of $\beta = 0.85$ to form the final context block. These ratios are adopted from the I-JEPA [12] framework, where they were shown to be empirically optimal. This two-step strategy ensures that there is no overlap between the context and the information it is trying to predict, which is crucial for a meaningful learning task.

3.1. Target

In this subsection, we explain how to form target blocks using the combined modality data obtained in the previous step. Target blocks (T_i) refer to the parts that need to be

predicted through the context block, effectively serving as label data. The fundamental idea of this predictive learning task is to force the model to learn high-level semantic representations by predicting the latent representation of a part of the input from another part. The method for constructing target blocks is as follows.

From the input $\mathbf{s}_{new}$, a new latent representation $\mathbf{s}_t$ is extracted using the target encoder f_t. This target encoder, f_t, acts as a "teacher" model, providing a stable and consistent semantic representation that serves as the prediction objective. By using a slowly evolving teacher, we prevent the model from collapsing to a trivial solution. From $\mathbf{s}_t$, the part corresponding to the video length N is extracted to form $\mathbf{s}_{vid}$. Target blocks are composed of a total of M blocks($T = \{\mathbf{t}_1, \cdots, \mathbf{t}_M\}$), where $\mathbf{t}$ refers to one target block. The sampling of target patches from $\mathbf{s}_{vid}$ follows the two-step procedure (with a ratio of $\alpha = 0.15$) described at the beginning of Section 3. It is important to note that sampling for target blocks is not done directly from $\mathbf{s}_{new}$ but from $\mathbf{s}_{vid}$ obtained through the target encoder. This ensures the prediction task focuses on the visual features, guided by the text query information already fused into $\mathbf{s}_{new}$. We do not apply any augmentation to the target blocks to ensure that the prediction goal remains a clean, undistorted representation of the original features. This detailed process ensures the accurate formation of target blocks, crucial for the subsequent steps in the representation learning framework.

All processes are repeated M times to form a total of T_i target blocks. Target blocks may overlap with each other. The indices of the patches used in the target blocks are stored separately for later use when designating context blocks. The target encoder is a teacher encoder, identical to the context encoder used for context blocks. It is a copy of the context encoder. The target encoder is not updated by backpropagation. Instead, the parameters are updated using an exponential moving average applied to the student encoder f_s, which is used later for the context blocks.

The use of a student-teacher architecture, where the context encoder acts as the student and the target encoder as the teacher, is crucial for preventing model collapse during representation learning. If a single encoder were used to generate both context and target representations, the model could easily learn a trivial solution by simply copying the input. The teacher encoder, which is updated slowly via an exponential moving average (EMA) of the student encoder's weights, provides a stable and consistent representation for the student to predict. This forces the student to learn higher-level semantic features rather than a simple identity function, a well-established technique to avoid collapse in self-supervised learning.

The overall process is illustrated in Equation (2).

$$\begin{aligned} &\mathbf{s}_t = f_t(\mathbf{s}_{new}) \\ &\mathbf{s}_{vid} = \mathbf{s}_t[:N] = \{s_{vid1}, \cdots, s_{vidN}\} \\ &\mathbf{t}_i = \text{RandomSample}(\mathbf{s}_{vid}, \alpha N) \\ &\mathbf{T} = \{\mathbf{T}_i, \cdots, \mathbf{T}_M\} \end{aligned} \tag{2}$$

3.2. Context

In this subsection, we explain how to form context blocks, which are used to predict the target blocks. The context block (C) provides the information given to the predictor during the prediction phase and is used to create the predicted target block for comparison with the actual target block.

The method used in previous research [12] involves extracting features directly from the original image without applying any augmentation. However, since our work uses pre-extracted features, this approach may impose some limitations on forming patches compared to methods that operate on raw data. Our proposed model applies slight aug-

mentations to the pre-extracted data to address these limitations. The augmentations include adding Gaussian noise and applying feature dropout. These augmentations are applied only when forming the context block, not the target block. This one-sided augmentation strategy is a core component of our approach. By corrupting the context, we create a more challenging prediction task, forcing the model to learn robust and generalizable features rather than simply memorizing the input data.

The method for forming context blocks is similar to the method for forming target blocks. From the input $\mathbf{s}_{new}$, only the part corresponding to the video is used (denoted as $\mathbf{c}_{vid}$). The context block is formed by taking the remaining parts of $\mathbf{c}_{vid}$ after excluding the patch indices used for the target blocks. From these remaining patches, the final context block is sampled with a rate of $\beta = 0.85$, as detailed in the procedure at the start of Section 3. Finally, the previously described augmentations are applied to the constructed context patches, and then the context encoder (f_c) is applied to form the context block (C). Unlike the target blocks, only one context block is formed, and this context block is used to predict M target blocks. The context encoder used here is the student encoder, which is updated through backpropagation. The context encoder is also used later in the VMR task as an encoder. The entire process is illustrated in Equation (3).

$$\begin{aligned} \mathbf{c}_{vid} &= \mathbf{s}_{new}[:N] = \{v_1, \cdots, v_N\} \\ C &= \text{RandomSample}(\mathbf{c}_{vid}, \beta N) \\ C &= f_c(Aug(C)) \end{aligned} \tag{3}$$

3.3. Latent Representation Prediction

This section describes the method for extracting M predicted target block representations using the context block (C) formed in the previous stage. Each predicted target block ($\hat{T}$) representation corresponds to a target block (T) that was established earlier. The method for predicting the target blocks is as follows. The input is C obtained from the context encoder and mask tokens (MASK). The number of mask tokens corresponds to the number of target blocks that need to be predicted. (MASK $= \{\text{mask}_1, \cdots, \text{mask}_J\}$) J denotes the number of target patches in one target block. Each mask token is a learnable parameter and has an added positional embedding. Context C and the mask tokens [MASK] are fed into the predictor (f_p) to obtain the predicted representation ($\hat{T}_i$) for each target block (T_i). The predictor's role is to synthesize the latent representation of the target by attending to the most relevant information within the context block C. Since predictions need to be made for M target blocks, this process is repeated M times.

In this stage, the predictor adopts the same DETR-style object-query decoder used in the downstream VMR head; we simply warm-start the VMR decoder with the weights learned during latent prediction [31]. While the encoders used in the previous stage are the same as those employed later in the VMR task, the decoder used here differs from the one used in the VMR task. The decoder is trained through backpropagation while representation learning. The overall process of latent representation prediction is illustrated in Equation (4):

$$\begin{aligned} \text{MASK} &= \{\text{mask}_i, \cdots, \text{mask}_J\} \\ \hat{T}_i &= f_p(C, \text{MASK}) \\ \hat{T} &= \{\hat{T}_i, \cdots, \hat{T}_M\} \end{aligned} \tag{4}$$

3.4. Loss Function

This section explains the loss function used in latent representation learning. The loss function, as shown in Equation (5), calculates the mean squared error (MSE) between

the latent representation of the target block (T) and the predicted latent representation of the target block ($\hat{T}$). We use MSE as it provides a direct and efficient way to minimize the Euclidean distance between the predicted and target vectors in the latent space. This encourages the predictor to generate representations that are not only semantically similar but also close in their geometric arrangement.

$$MSE = \frac{1}{M}\sum_{i=1}^{M}(\hat{T}_i - T_i) \tag{5}$$

The calculated loss is used to update the predictor and context encoder parameters through backpropagation. The parameters of the target encoder are not updated through backpropagation of the loss; instead, they are updated using the exponential moving average of the context encoder's parameters [15,32].

3.5. Video Moment Retrieval

After completing the first stage of latent representation learning, the model proceeds to the main task, that is, video moment retrieval (VMR). VMR involves identifying the segment of a video that corresponds to a given text query. The general framework for this stage, based on the QD-DETR [2] model, is illustrated in Figure 3.

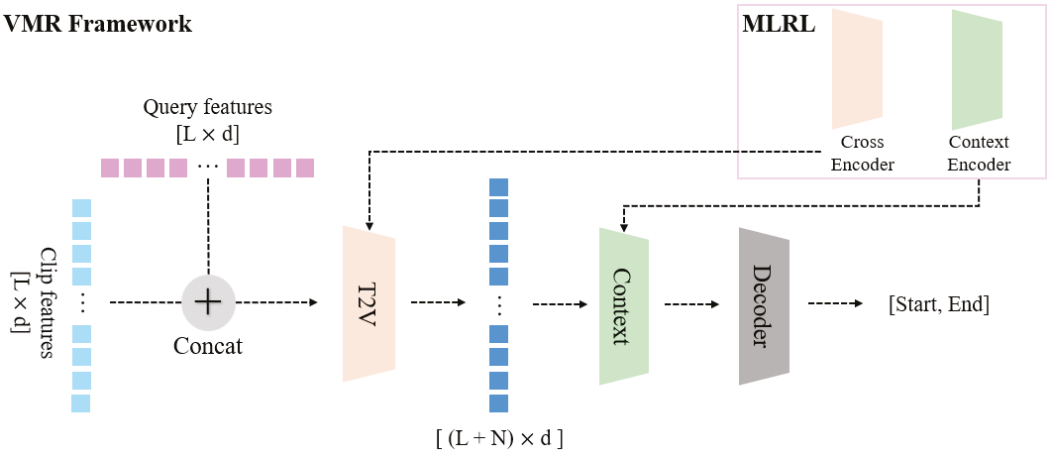


Figure 3. Video moment retrieval framework. The video moment retrieval (VMR) framework, based on QD-DETR [2]. Our proposed MLRL module (top right) replaces the original encoders.

The input consists of video clip features and text queries. Using data from both modalities, the model ultimately predicts the relevant moment in the video as an object detection output, specifying the middle timestamp and the span on either side. Additionally, QD-DETR [2] performs highlight detection using the saliency scores provided by the QVHighlight dataset [1].

The method proposed in this paper aims to verify the performance changes in downstream tasks when additional representation learning is applied to pre-extracted feature data. Therefore, the training method for the VMR stage remains unchanged. As shown in Figure 3, the key difference in our approach is that the original T2V encoder and main encoder of the QD-DETR [2] model are replaced with the cross-encoder and context encoder learned during the MLRL stage.

4. Results

4.1. Dataset and Metrics

This paper uses the QVHighlight [1] dataset, which provides labeled data for video moment retrieval and highlight detection tasks. The dataset consists of 10,148 videos

sourced from YouTube, accompanied by 10,310 human-written text queries. Following the official splits, the dataset is divided into 7088 videos for the training set, 1518 for the validation set, and 1542 for the test set.

Each video is approximately 150 s long, and the queries average 11.3 words. For our experiments, we use the pre-extracted features provided by the original authors. Specifically, the video features are 2304-dimensional vectors extracted from a pre-trained SlowFast network, and the query features are 768-dimensional vectors from a pre-trained RoBERTa model. The original video data is divided into 2 s clips for this feature extraction. Furthermore, the dataset provides saliency scores for each clip, indicating its relevance to the query, which is used for the highlight detection task.

To validate the generalizability of our approach, we also evaluate its performance on the Charades-STA dataset [6], a widely-used benchmark for sentence-based temporal activity localization. Derived from the original Charades dataset, it consists of 12,408 temporal annotations across 6672 training videos and 3720 annotations for 1334 testing videos. For fair performance analysis, we used the same metrics as the baseline QD-DETR [2] to measure performance. We measured performance using recall and mean average precision (mAP). Recall was measured at IoU thresholds of 0.5 and 0.7; mAP was measured at IoU values of 0.5, 0.7, and as the average over multiple thresholds in the range [0.5:0.95; interval = 0.05].

4.2. Experiment Setup

The proposed method in this paper, MLRL, is divided into two stages. One is the latent representation learning stage, where additional representation learning is conducted on the pre-extracted feature data proposed in this paper. The other is the video moment retrieval stage, where the downstream task is performed. This paper proposes a methodology for the latent representation learning stage and, consequently, does not significantly modify the moment retrieval stage. Therefore, most settings are the same as the baseline QD-DETR [2]. In this paper, experiments were conducted on two scenarios. One scenario involves conducting representation learning and moment retrieval separately, and the other consists of conducting both simultaneously. In the simultaneous scenario, the settings for representation learning are the same as QD-DETR [2]. In contrast, in the separate scenario, the settings are almost identical to those in I-JEPA [12]. The differences are that the batch size is set to 32 and the epochs are set to 100. The separate regime reflects a practical use case in which the encoder is refined once on a large, unlabeled corpus and then reused across multiple retrieval heads without rerunning the expensive detector, thereby reducing GPU-hours and memory peaks. The main settings in the latent representation learning stage are as follows. The hyperparameters for sampling, namely the target block ratio α and the context block ratio β, were set to 0.15 and 0.85, respectively. This decision is based on the optimal values presented in the original I-JEPA [12] paper, where extensive ablation studies (Assran et al., 2023 [12]) confirmed that this configuration provides the best performance by balancing the complexity of the prediction task and the informativeness of the context. The weights of the target encoder are the same as those of the context encoder. At this time, the target encoder is updated not through backpropagation but through an exponential moving average (EMA). The momentum of the EMA increases linearly from 0.996 to 1.0 during the training process. All experiments were conducted in an environment equipped with two Intel(R) Xeon(R) Gold 6248R CPUs, 345GiB of RAM, and eight NVIDIA TITAN RTX GPUs (each with 24 GB VRAM).

4.3. Experiment Result

This section introduces the experimental results for MLRL. The experiments compare the performance of MLRL with the baseline [2] and the original version of [1,2]. Table 1 shows the experimental results.

Before explaining the experimental results, let us first define the terms used in the table. The mAP in the table is the average value of IoU thresholds from 0.5 to 0.95. R represents the recall value. Extra training data indicates whether other datasets were used in addition to [1] during the training process. Rep/MR means that latent representation learning and moment retrieval were conducted separately. This means the encoders were trained through representation learning first and then used in moment retrieval. Rep+MR means that latent representation learning and moment retrieval were conducted simultaneously. The final loss is calculated by combining the loss of the representation learning method and the loss of the moment retrieval method within one framework.

Comparing the results with the baseline QD-DETR(val) in Table 1, our method shows clear performance gains. The Rep/MR model, for instance, achieved an increase of +1.05 in mAP and +0.89 in R1@0.7. This demonstrates that enhancing feature representation through our proposed MLRL is more effective than simply using pre-extracted features for the downstream task.

Table 1. Comparison results of the multimodal latent representation learning (MLRL). This table compares the performance of MLRL with previous models, Moment-DETR [1] and QD-DETR [2]. The term (test) refers to evaluations conducted on the test dataset of QVHighlight [1], while (val) indicates evaluations performed on the validation dataset. The performance of MLRL was assessed using the validation set of the QVHighlight [1] dataset.

Model	mAP	mAP@0.5	mAP@0.75	R1@0.5	R1@0.7	Extra Data
Moment-DETR [1]	30.73	54.82	29.40	52.89	33.02	x
QD-DETR(w/audio) [2]	40.19	63.04	40.10	63.06	45.10	x
QD-DETR(test) [2]	39.86	62.52	40.10	62.40	45.10	x
QD-DETR(val) [2]	41.22	62.23	41.82	62.68	46.66	x
MLRL(Rep/MR)	**42.27**	62.20	**42.54**	62.26	47.55	x
MLRL(Rep+MR)	42.06	**62.32**	42.06	**62.32**	**47.65**	x

To provide a broader context for our work and address the need for additional benchmarks, we also compare our method with recent state-of-the-art models that leverage large language models (LLMs). The results are presented in Table 2. As shown in Table 2, recent state-of-the-art models leveraging large language models (LLMs) demonstrate high performance on the QVHighlight benchmark. However, a direct comparison of raw performance metrics with our MLRL framework requires careful consideration of the models' scale and design philosophy. These LLM-based models [23,24,30,33], such as LLaVA-MR [23] and InternVideo2-6B [24], often contain billions of trainable parameters and demand substantial computational resources for training and inference. In contrast, the primary goal of our MLRL is not to compete with these massive models in raw performance, but to provide a lightweight and efficient method for significantly enhancing pre-extracted features with minimal overhead. Therefore, the results in Table 2 should be interpreted from two perspectives, as follows: while LLMs define the upper bound of performance on this benchmark, our MLRL framework demonstrates a more practical and resource-efficient pathway to achieving substantial improvements over strong, non-LLM baselines like QD-DETR [2]. This highlights the value of our approach for scenarios where computational efficiency is a critical consideration.

To validate its generalizability, we also evaluated our framework on the Charades-STA [6] dataset, with results in Table 3. On this dataset, our MLRL(Rep+MR) model

improves upon the QD-DETR baseline by +0.70 points on R1@0.5 and a more significant +1.99 points on R1@0.7. These results confirm that the benefits of MLRL extend to other standard benchmarks, proving its robustness.

Table 2. Performance comparison with LLM-based models on the QVHighlight dataset. This table compares our MLRL models against recent LLM-based methods [23,24,30,33]. All models are evaluated on the QVHighlight validation set without external training data. The best performance for each metric is underlined.

Model	mAP	mAP@0.5	mAP@0.75	R1@0.5	R1@0.7
UniVTG (w/ PT)	43.63	64.06	45.02	65.43	50.06
InternVideo2-6B	49.24	-	-	71.42	56.45
FlashVTG	52.00	<u>72.33</u>	53.85	70.69	53.96
LLaVA-MR	<u>52.73</u>	69.41	<u>54.40</u>	<u>76.59</u>	<u>61.48</u>
MLRL(Rep/MR)	**42.27**	62.20	**42.54**	62.26	47.55
MLRL(Rep+MR)	42.06	**62.32**	42.06	**62.32**	**47.65**

Table 3. Performance comparison on the Charades-STA dataset. Performance on the Charades-STA [6] dataset to evaluate generalizability. 'w/ PT' indicates the Moment-DETR [1] model was pre-trained on 10K HowTo100M videos.

Model	R1@0.5	R1@0.7
Moment-DETR [1]	53.63	31.37
Moment-DETR [1] w/ PT (on 10K HowTo100M videos)	55.65	34.17
QD-DETR [2] (Only Video)	57.31	32.55
MLRL(Rep/MR)	57.96	**33.77**
MLRL(Rep+MR)	**58.01**	33.65

4.4. Ablation Study

In this section, we compare the results based on different modalities of data used for representation learning through the proposed MLRL. The comparison is conducted across four modalities, as follows: Clip/Query, Clip, Query, and None. 'Clip/Query' refers to the case where both clip features and query features are utilized together, 'Clip' refers to the case where only clip features are used, 'Query' refers to the case where only query features are used, and 'None' represents the baseline where latent representation learning is not applied. The cross-encoder, which combines the two modalities, is not applied to the Clip and Query cases. In the case of Query, since there is no clip, a mask is added to the tokens during context and target sampling.

As shown in Table 4, the results show that using Clip/Query or Clip alone leads to an improvement in metrics. Specifically, for Clip/Query, there was an increase of +1.05 in mAP, +0.72 in mAP@0.75, and +0.89 in R1@0.7. For Clip, the increases were +0.95 in mAP, +0.23 in mAP@0.5, +0.26 in mAP@0.75, and +0.65 in R1@0.7. On the other hand, when using only Query, the metrics tend to be similar to or even lower than when no features were used at all.

Table 4. Ablation study of the latent representation learning. This table presents the results when applying latent representation learning to different modalities. 'Clip/Query' shows the results when both clip features and query features are used together, 'Clip' shows the result when only clip features are used, and 'None' represents the results without applying latent representation learning.

Modality	mAP	mAP@0.5	mAP@0.75	R1@0.5	R1@0.7
Clip/Query	42.27	62.20	42.54	62.26	47.55
Clip	42.17	62.46	42.08	62.35	47.31
Query	41.03	61.97	41.65	62.61	46.23
None(Baseline)	41.22	62.23	41.82	62.68	46.66

In summary, the results indicate that using Clip/Query or Clip leads to a noticeable improvement in performance, highlighting the critical role of visual information (clip features) in video moment retrieval tasks. When both clip and query features are used together, performance improves at higher thresholds, suggesting that combining visual and textual information can have a complementary effect in enhancing model accuracy. On the other hand, when only query features are used, performance does not improve significantly or may even decline, indicating that textual information alone may not be sufficient for this task. Overall, these findings emphasize the importance of integrating visual information and multimodal data to achieve more robust and accurate results in video moment retrieval.

4.5. Qualitative Result

In this section, we examine how the performance of the proposed MLRL model is reflected in actual qualitative results. Figure 4 shows the ground truth (G.T.), the predicted segments from MLRL, and the predicted segments from QD-DETR [2] for a portion of the test. Upon review, it is evident that both QD-DETR [2] and MLRL predict segments similarly to the G.T. The key difference is that while QD-DETR [2] separates the segments like the G.T., MLRL predicts a single, combined segment. However, upon closer inspection of the G.T., we find that the portion where the segments are split, which corresponds to the part not related to the query (a close-up of a woman's face), lasts for only about one second and is semantically part of a single segment. In this sense, it can be concluded that MLRL, through additional training that enhances the video's representation, is capable of segment prediction that considers the semantic aspects of the video.

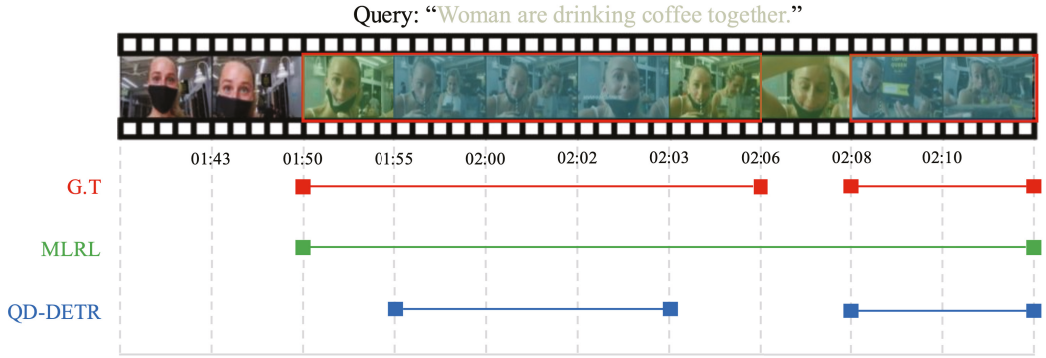


Figure 4. Qualitative result. This image shows the results of a qualitative experiment comparing MLRL with ground truth (G.T.) and QD-DETR.

5. Discussion

In this section, we discuss the issues discovered during the experiments with the model and potential future research directions. One problem identified during the experiments

is the ambiguous utilization of query features in the representation learning stage. In the current method, query features are used by concatenating clip features and query features and then inputting them into a cross-encoder to transfer the information of query features to clip features. We believe that using a method where the query directly influences the loss calculation in representation learning can result in better performance in the multimodal task of moment retrieval.

Based on the current research, we introduce potential future work. Research could explore ways to make query features have a more direct impact on representation learning, rather than just using them in cross-attention. For instance, a method could be proposed in which context blocks and target blocks are selected based on their relevance to the query.

Another direction for future research could be to demonstrate the agnostic nature of the model. The method proposed in this study pertains to representation learning and is an agnostic approach that can be applied not only to the video moment retrieval (VMR) method used in this study, but also to other VMR methods [3–5]. Therefore, future research could involve applying the proposed method to other VMR methods to verify that the method is indeed model agnostic.

6. Conclusions

This paper addresses the challenge of processing large-scale video sensor data, where direct feature extraction from raw signals is a major computational bottleneck. We propose the multimodal latent representation learning (MLRL) framework, which enhances the expressiveness of pre-extracted features through an additional representation learning stage. Experiments confirmed that our model significantly outperforms approaches that simply use the pre-extracted features for downstream tasks. This result indicates that MLRL offers a practical pathway to more accurately interpret sensor data without costly reprocessing, thereby contributing to the development of more efficient and powerful sensor systems. While the effectiveness of MLRL has been demonstrated, there remains room for future research, such as improving the utilization of query features and proving the model's agnostic nature across different architectures.

Author Contributions: Conceptualization, J.H. and J.K.; methodology, J.H. and J.K.; software, J.H.; validation, M.J.; formal analysis, M.J. and J.K.; investigation, J.K.; resources, J.H. and J.K.; data curation, J.H. and J.K.; writing—original draft preparation, J.H., M.J. and J.K.; writing—review and editing, J.H., M.J. and J.K.; visualization, J.K.; supervision, J.K.; project administration, J.K.; funding acquisition, J.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [RS-2021-II211341, Artificial Intelligence Graduate School Program (Chung-Ang University)] and supported by the Chung-Ang University Research Grants in 2024.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in https://github.com/jayleicn/moment_detr, accessed on 15 January 2025.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Lei, J.; Berg, T.L.; Bansal, M. Detecting moments and highlights in videos via natural language queries. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 11846–11858.

2. Moon, W.; Hyun, S.; Park, S.; Park, D.; Heo, J.P. Query-dependent video representation for moment retrieval and highlight detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 23023–23033.
3. Moon, W.; Hyun, S.; Lee, S.; Heo, J.P. Correlation-guided query-dependency calibration in video representation learning for temporal grounding. *arXiv* **2023**, arXiv:2311.08835.
4. Lee, P.; Byun, H. BAM-DETR: Boundary-Aligned Moment Detection Transformer for Temporal Sentence Grounding in Videos. *arXiv* **2023**, arXiv:2312.00083.
5. Xiao, Y.; Luo, Z.; Liu, Y.; Ma, Y.; Bian, H.; Ji, Y.; Yang, Y.; Li, X. Bridging the gap: A unified video comprehension framework for moment retrieval and highlight detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–18 June 2024; pp. 18709–18719.
6. Gao, J.; Sun, C.; Yang, Z.; Nevatia, R. Tall: Temporal activity localization via language query. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5267–5275.
7. Escorcia, V.; Soldan, M.; Sivic, J.; Ghanem, B.; Russell, B. Temporal localization of moments in video collections with natural language. *arXiv* **2019**, arXiv:1907.12763.
8. Wang, J.; Ma, L.; Jiang, W. Temporally grounding language queries in videos by contextual boundary-aware prediction. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12168–12175.
9. Zhang, S.; Peng, H.; Fu, J.; Luo, J. Learning 2d temporal adjacent networks for moment localization with natural language. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12870–12877.
10. Gao, J.; Xu, C. Fast video moment retrieval. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 10–17 October 2021; pp. 1523–1532.
11. LeCun, Y. A Path Towards Autonomous Machine Intelligence Version 0.9.2, 2022-06-27. 2022. Available online: https://openreview.net/forum?id=BZ5a1r-kVsf (accessed on 20 January 2025).
12. Assran, M.; Duval, Q.; Misra, I.; Bojanowski, P.; Vincent, P.; Rabbat, M.; LeCun, Y.; Ballas, N. Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture. *arXiv* **2023**, arXiv:2301.08243. [CrossRef]
13. Baevski, A.; Hsu, W.N.; Xu, Q.; Babu, A.; Gu, J.; Auli, M. data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language. *arXiv* **2022**, arXiv:2202.03555. [CrossRef]
14. Bardes, A.; Garrido, Q.; Ponce, J.; Chen, X.; Rabbat, M.; LeCun, Y.; Assran, M.; Ballas, N. Revisiting Feature Prediction for Learning Visual Representations from Video. *arXiv* **2024**, arXiv:2404.08471. [CrossRef]
15. Chen, X.; Xie, S.; He, K. An empirical study of training self-supervised vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 10–17 October 2021; pp. 9640–9649.
16. Chen, T.; Kornblith, S.; Swersky, K.; Norouzi, M.; Hinton, G.E. Big self-supervised models are strong semi-supervised learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 22243–22255.
17. Chen, X.; He, K. Exploring simple siamese representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 15750–15758.
18. Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; Joulin, A. Unsupervised learning of visual features by contrasting cluster assignments. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 9912–9924.
19. Qian, R.; Meng, T.; Gong, B.; Yang, M.H.; Wang, H.; Belongie, S.; Cui, Y. Spatiotemporal contrastive video representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 6964–6974.
20. Kuang, H.; Zhu, Y.; Zhang, Z.; Li, X.; Tighe, J.; Schwertfeger, S.; Stachniss, C.; Li, M. Video contrastive learning with global context. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 10–17 October 2021; pp. 3195–3204.
21. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2019**, arXiv:1810.04805. [CrossRef]
22. He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; Girshick, R. Masked Autoencoders Are Scalable Vision Learners. *arXiv* **2021**, arXiv:2111.06377. [CrossRef]
23. Lu, W.; Jia, Z.; Xie, C.W.; Fan, L.W.; Xiang, T.Z.; Luo, D.S. LLaVA-MR: Large Language-and-Vision Assistant for Video Moment Retrieval. *arXiv* **2024**, arXiv:2411.14505.
24. Wang, Y.; Li, K.; Li, X.; Yu, J.; He, Y.; Wang, C.; Chen, G.; Pei, B.; Yan, Z.; Zheng, R.; et al. InternVideo2: Scaling Foundation Models for Multimodal Video Understanding. *arXiv* **2024**, arXiv:2403.15377. [CrossRef]
25. Assran, M.; Balestriero, R.; Duval, Q.; Bordes, F.; Misra, I.; Bojanowski, P.; Vincent, P.; Rabbat, M.; Ballas, N. The Hidden Uniform Cluster Prior in Self-Supervised Learning. *arXiv* **2022**, arXiv:2210.07277. [CrossRef]
26. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 213–229.

27. Cheng, B.; Misra, I.; Schwing, A.G.; Kirillov, A.; Girdhar, R. Masked-attention mask transformer for universal image segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 1290–1299.
28. Cheng, B.; Schwing, A.; Kirillov, A. Per-pixel classification is not all you need for semantic segmentation. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 17864–17875.
29. Huang, K.C.; Wu, T.H.; Su, H.T.; Hsu, W.H. Monodtr: Monocular 3d object detection with depth-aware transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 4012–4021.
30. Lin, K.Q.; Zhang, P.; Chen, J.; Pramanick, S.; Gao, D.; Wang, A.J.; Yan, R.; Shou, M.Z. UniVTG: Towards Unified Video-Language Temporal Grounding. *arXiv* **2023**, arXiv:2307.16715. [CrossRef]
31. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2023**, arXiv:1706.03762. [CrossRef] [PubMed]
32. Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; Joulin, A. Emerging Properties in Self-Supervised Vision Transformers. *arXiv* **2021**, arXiv:2104.14294. [CrossRef]
33. Cao, Z.; Zhuang, J.X.; Li, Z.Y.; Zhang, X.Y.; Wang, L.M.; Jiang, Y.G. FlashVTG: Feature Layering and Adaptive Score Handling Network for Video Temporal Grounding. *arXiv* **2024**, arXiv:2412.13441. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

sensors

MDPI

Article

Toward a Recognition System for Mexican Sign Language: Arm Movement Detection

Gabriela Hilario-Acuapan, Keny Ordaz-Hernández *, Mario Castelán and Ismael Lopez-Juarez *

Robotics and Advanced Manufacturing Department, Centre for Research and Advanced Studies (CINVESTAV), Ramos Arizpe 25900, Mexico; gabriela.hilario@cinvestav.mx (G.H.-A.); mario.castelan@cinvestav.mx (M.C.)
* Correspondence: keny.ordaz@cinvestav.edu.mx (K.O.-H.); ismael.lopez@cinvestav.edu.mx (I.L.-J.)

Abstract: This paper describes ongoing work surrounding the creation of a recognition system for Mexican Sign Language (LSM). We propose a general sign decomposition that is divided into three parts, i.e., hand configuration (HC), arm movement (AM), and non-hand gestures (NHGs). This paper focuses on the AM features and reports the approach created to analyze visual patterns in arm joint movements (wrists, shoulders, and elbows). For this research, a proprietary dataset—one that does not limit the recognition of arm movements—was developed, with active participation from the deaf community and LSM experts. We analyzed two case studies involving three sign subsets. For each sign, the pose was extracted to generate shapes of the joint paths during the arm movements and fed to a CNN classifier. YOLOv8 was used for pose estimation and visual pattern classification purposes. The proposed approach, based on pose estimation, shows promising results for constructing CNN models to classify a wide range of signs.

Keywords: Mexican Sign Language; dynamic signs; pattern analysis; pose-based approach; computer vision; machine learning; CNN; YOLOv8; arm movement

1. Introduction

Deafness or hearing loss is the partial or total loss of the ability to hear sounds in one or both ears. The World Health Organization's most recent World Hearing Report [1] estimates that more than 1.5 billion people have some degree of hearing loss. Approximately 430 million of them have moderate or greater hearing loss in their better ear; this number is expected to increase to 700 million people by 2050.

According to the Ministry of Health [2], approximately 2.3 million people in Mexico have hearing disabilities. This vulnerable group faces significant levels of discrimination and limited employment opportunities. Additionally, this health condition restricts access to education, healthcare, and legal services, further exacerbating social inequalities and limiting opportunities for integration. One of the primary challenges faced by the deaf community is communication with hearing individuals, as linguistic differences hinder social and workplace interactions. While technology has proven useful in reducing some of these barriers, deaf individuals often rely on the same technological tools as the hearing population, such as email and text messaging applications. However, these tools are not always effective, as not all deaf individuals are proficient in written Spanish.

In the Americas, the most widely studied sign languages are American Sign Language (ASL) and Brazilian Sign Language (LIBRAS), which have facilitated research and technological advancements aimed at improving communication with the deaf community. An example of one innovation is SLAIT [3], a startup that emerged from a research project at

Aachen University of Applied Sciences in Germany. During this research, an ASL recognition engine was developed using MediaPipe and recurrent neural networks (RNNs). Similarly, Ref. [4] announced an innovative project in Brazil that uses computer vision and artificial intelligence to translate LIBRAS into text and speech in real time. Although this technology is still undergoing internal testing, the developers claim that after four years of work, the system has reached a significant level of maturity. This technology was developed by Lenovo researchers in collaboration with the Center for Advanced Studies and Systems in Recife (CESAR), which has already patented part of this technology [5]. The system is capable of recognizing the positions of arm joints, fingers, and specific points on the face, similar to SLAIT. From this data, it processes facial movements and gestures, enabling the identification of sentence flow and the conversion of sign language into text. CESAR and Lenovo believe that their system has the potential to become a universally applicable tool.

Compared to speech recognition and text translation systems, applications dedicated to sign language (SL) translation remain scarce. This is partly due to the relatively new nature of the field and the inherent complexity of sign language recognition (SLR), which involves visual, spatial, and gestural elements. Recognizing sign language presents a significant challenge, primarily due to limited research and funding. This highlights the importance of promoting research into the development of digital solutions that enhance the quality of life for the deaf community (c.f. [6]). However, researchers agree that the key factor for developing successful machine learning models is data (c.f. [7]). In this regard, for SLs like LSM, existing databases are often inadequate in terms of both size and quality, hindering the advancement of these technologies. Also, sensing technology has a fundamental role in the reliability of incoming data. This is the main reason why SLR is broadly divided into two branches, i.e., contact sensing and contactless sensing.

Sign data acquisition with contact relies on gloves [8], armbands [9], wearable inertial sensors [10,11], or electromyographic (EMG) signals [12]. In contrast, contactless sign data acquisition is mainly divided into two types, depending on the kind of hardware, that is, simple hardware (color or infrared cameras) or specialized hardware (e.g., depth sensors, optical 3D sensors [13], commercial WiFi devices [14], and ultrasonic devices [15]).

This classification is similar to the one presented by [16] (Figure 1), except that their sign data acquisition approaches are divided into sensor-based approaches and vision-based approaches. We present several examples of sign language research and related work, along with various approaches to sign data acquisition, as detailed in Table 1.

In Table 1, we include information regarding the features of signs that are included in the sign data acquisition for each reported work. Instead of using the separation employed by [17] (facial, body, and hand features), we propose our own decomposition into hand configurations (HCs), arm movements (AMs), and non-hand gestures (NHGs); see Figure 1. This is a fundamental concept of our research, so this decomposition is discussed in more detail in Section 1.1.2. The separation of facial, body, and hand features is a concept commonly seen in pose estimators (such as MediaPipe [18]), which are also common in SL research, as presented in Table 1. It is also possible to observe that most SL research is focused on the HC features.

We will now present the scientific context of LSM research. First, we present the known datasets and then present studies about LSM recognition and analysis.

LSM is composed of two parts, i.e., dactylology (fingerspelling) and ideograms ([19], p. 12). Dactylology is a small subset of LSM and basically consists of letters of the alphabet, where most parts are static signs. A few signs for numbers are also static. Due to the small, nevertheless important, role of dactylology, we are interested in LSM ideogram datasets. To the best of our knowledge, there are three publicly available ideogram-focused datasets.

Two of them are visual, i.e., (i) the MX-ITESO-100 preview [20], which contains video clips of 11 signs from 3 signers (out of 100 signs, although not all are currently available), and (ii) the Mexican Sign Language dataset [21,22], which includes image sequences of 249 signs from 11 signers. The third dataset, consisting of keypoints, is provided by [23]; this dataset contains 3000 samples of 30 signs from 4 signers (8 letters, 20 words, and 2 phrases). This was constructed by processing the RGBD data into keypoints by means of the MediaPipe [18] tool, but the unprocessed visual data is not provided. A comparison of these datasets, along with LSM glossaries, is provided in Tables 2 and 3.

Table 1. Sign language research and related work.

Ref.	SL	Sign Group *	Sign Type	Sign Features †	Sensor/Tool
Yao et al. (2025) [24]	ASL	L, N, P	Static	HC	Hydrogel strain sensor
Chiradeja et al. (2025) [8]	-	S	Dynamic	HC	Gloves
Rodríguez-Tapia et al. (2019) [10]	ASL	W	Dynamic	HC	Myoelectric bracelets
Filipowska et al. (2024) [12]	PJM	W	Dynamic	HC	EMG
Umut and Kumdereli (2024) [9]	TSL	W	Dynamic	HC, AM	Myo armbands (IMU + sEMG)
Gu et al. (2024) [11]	ASL	W, S	Dynamic	HC, AM	IMUs
Wei et al. (2025) [25]	-	W	Dynamic	HC	Gloves
Wang et al. (2025) [26]	ASL	L	Both	HC	Triboelectric sensor
Urrea et al. (2023) [27]	ASL	L, W	Static	HC	Camera/MediaPipe
Al-Saidi et al. (2024) [16]	ArSL	L	Static	HC	Camera/MediaPipe
Niu (2025) [28]	ASL	L	Static	HC	Camera
Hao et al. (2020) [14]	-	W	Dynamic	HC	WiFi
Galván-Ruiz et al. (2023) [13]	LSE	W	Dynamic	HC	Leap motion
Wang et al. (2023) [15]	CSL	W, P	Dynamic	HC	Ultrasonic
Raihan et al. (2024) [29]	BdSL	L, N, W, P	Dynamic	HC	Kinect
Woods and Rana (2023) [30]	ASL	W	Dynamic	AM, NHG	Camera/OpenPose
Eunice et al. (2023) [31]	ASL	W	Dynamic	HC, AM, NHG	Camera/Sign2Pose, YOLOv3
Gao et al. (2024) [17]	ASL, TSL	W	Dynamic	HC, AM, NHG	Camera, Kinect
Kim and Baek (2023) [32]	DGS, KSL	W, S	Dynamic	HC, AM, NHG	Camera/AlphaPose
Boháček and Hrúz (2022) [33]	ASL, LSA	W	Dynamic	HC, AM, NHG	Camera/Vision API (Apple)
Cihan Camgöz et al. (2020) [34]	DGS	S	Dynamic	HC, AM, NHG	Camera
Miah et al. (2024) [35]	ASL, PSL, LSM	L, W, P	Dynamic	HC, AM, NHG	Camera/MediaPipe, OpenPose
Gil-Martín et al. (2023) [36]	LSE	L, N, W	Both	HC, AM, NHG	Virtual ‡ camera/MediaPipe
Villa-Monedero et al. (2023) [37]	LSE	L, N, W	Both	HC, AM, NHG	Virtual ‡ camera/MediaPipe
Current study	LSM	W, P	Dynamic	AM	Camera/YOLOv8

* L: alphabet letter; N: number; W: word; P: phrase; S: sentence; † HC: hand configuration; AM: arm movement; NHG: non-hand gesture. ‡ Virtual camera since the dataset was created using synthetic avatars. SL names are provided in the Abbreviations section. Top part: Sign data acquisition with contact sensing. Bottom part: Contactless sign data acquisition.

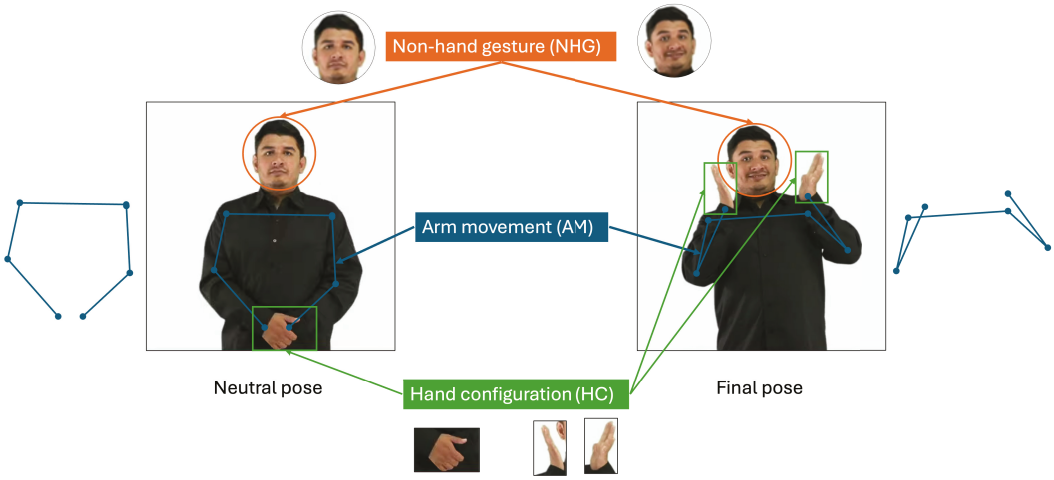


Figure 1. Sign features: hand configuration (HC), arm movement (AM), and non-hand gesture (NHG). "Surprise!" sign images were taken from screenshots of the corresponding YouTube video of the GDLSM [38]; see Appendix A.

Table 2. LSM datasets and glossaries.

Ref.	Type	Sign Group *	Sign Signal	Samples
DIELSEME 1 (2004) [39]	Glossary †	535 W	Visual	1 video per sign
DIELSEME 2 (2009) [40]	Glossary †	285 W	Visual	1 video per sign
GDLSM (2024) [38]	Glossary	27 L, 49 N, 667 W, 4 P	Visual	1 video per sign ‡
MX-ITESO-100 (2023) [20]	Dataset	96 W, 4 P	Visual	50 videos per sign
Mexican Sign Language dataset (2024) [22]	Dataset	243 W, 6 P	Visual	11 image sequences per sign
Mexican Sign Language Recognition (2022) [23]	Dataset	8 L, 21 W, 1 P	Keypoints	100 samples per sign

* L: alphabet letter; N: number; W: word; P: phrase. † According to [41], DIELSEME 1 and 2 are actually glossaries and not dictionaries. The three LSM glossaries contain only one sample per sign, whereas the datasets include multiple samples per sign. ‡ Their site reports 719 videos, but only 715 were found; also, the 32 videos in the "*Estados y capitales*" thematic category include 2 signs per video.

Table 3. LSM datasets and glossaries: sign and signal properties.

Ref.	Sign Features	Signal Properties	File Format	Comments
DIELSEME 1 (2004) [39]	HC, AM *, NHG	320 × 234 @ 12 fps	SWF videos	
DIELSEME 2 (2009) [40]	HC, AM, NHG	720 × 405 @ 30 fps	FLV videos	
GDLSM (2024) [38]	HC, AM, NHG	1920 × 1080 @ 60 fps	videos	Hosted on a streaming platform; c.f. Appendix A
MX-ITESO-100 (2023) [20]	HC, AM, NHG	512 × 512 @ 30 fps	MP4 videos	Preview only ‡
Mexican Sign Language dataset (2024) [22]	HC, AM *	640 × 480	JPEG images	Blurred faces
Mexican Sign Language Recognition (2022) [23]	HC, AM, NHG	20 × 201 array	CSV files	One row per frame, 67 (x, y, z) keypoints

* In those cases, the background and clothing are black, so the segmentation of skin (hand and face) is easier, but tracking joints for AM is more difficult. ‡ Only 11 signs (words) are available in the public preview. Also, the 50 samples of each sign were performed by a single subject.

Regarding LSM studies, most of the SLR research on LSM mainly focuses on classifying static letters and numbers using classical machine learning techniques and convolutional neural networks (CNNs) [42–49]. Using the classification provided by [16], there are four classes of signs: (i) continuous signs, (ii) isolated signs, (iii) letter signs, and (iv) number signs. In LSM, most of the signs in the last three categories are static signs. But signing in LSM is generally highly dynamic and continuous, since most signs are ideograms, as mentioned before.

In terms of dynamic sign recognition, early studies focused on classifying letters and numbers with motion. For example, Ref. [50] used the CamShift algorithm to track the hand trajectory, generating a bitmap that captures the pixels of the hand path; these bitmaps were then classified using a CNN. Another approach, presented in [51], involved obtaining the coordinates (x,y) of 22 keypoints of the hand using an Intel RealSense sensor, which were used as training data for a multilayer perceptron (MLP) neural network. Finally, in [52], 3D body cue points obtained with MediaPipe were used to train two recurrent neural networks (RNNs), i.e., LSTM and GRU.

In more recent research, in addition to letters and numbers, some simple words and phrases were included. Studies such as Refs. [53–55] used MLP-type neural networks, while others, such as Ref. [23], used more advanced RNN models. In Ref. [20], CNNs were used to extract features from the frames of a series of videos, which were then used as input to an LSTM model.

On the other hand, Ref. [56] presented a method for dynamic sign classification that involves extracting a sequence of frames, followed by a segmentation process using neural networks based on color, resulting in the segmentation of the skin of the hands and face. To classify the signs, four classical machine learning algorithms are compared, i.e., Bayesian classifier, decision trees, SVM, and NN.

Although research on LSM recognition has been conducted for several years, progress in this area has been slow and limited compared to other SLs. A common approach is to use computer vision techniques such as CNNs to build automatic sign recognition systems. However, with the recent emergence of pose recognition models, such as MediaPipe and YOLOv8, there is a trend in both LSM and other sign languages to use these tools to train more complex models, such as RNNs, or more sophisticated architectures, such as Transformers. A comparison of the studies mentioned here, with additional details, is shown in Table 4.

Table 4. LSM research.

Ref.	Sign Group *	Sign Type	Sign Feature	Sensor/Tool
Solís et al. (2016) [42]	L	Static	HC	Camera
Carmona-Arroyo et al. (2021) [43]	L	Static	HC	Leap Motion, Kinect
Salinas-Medina and Neme-Castillo (2021) [44]	L	Static	HC	Camera
Rios-Figueroa et al. (2022) [45]	L	Static	HC	Kinect
Morfín-Chávez et al. (2023) [46]	L	Static	HC	Camera/MediaPipe
Sánchez-Vicinaiz et al. (2024) [47]	L	Static	HC	Camera/MediaPipe
García-Gil et al. (2024) [48]	L	Static	HC	Camera/MediaPipe
Jimenez et al. (2017) [49]	L, N	Static	HC	Kinect
Martínez-Gutiérrez et al. (2019) [51]	L	Both	HC	RealSense f200
Rodriguez et al. (2023) [52]	L, N	Both	HC	Camera/MediaPipe
Rodriguez et al. (2025) [57]	L, N	Both	HC	Camera/MediaPipe
Martinez-Seis et al. (2019) [50]	L	Both	AM	Camera
Mejía-Peréz et al. (2022) [23]	L, W	Both	HC, AM, NHG	OAK-D/MediaPipe
Sosa-Jiménez et al. (2022) [58]	L, N, W	Both	HC, body but not NHG	Kinect
Sosa-Jiménez et al. (2017) [53]	W, P	Dynamic	HC, AM	Kinect/Pose extraction
Varela-Santos et al. (2021) [59]	W	Dynamic	HC	Gloves
Espejel-Cabrera et al. (2021) [56]	W, P	Dynamic	HC	Camera
García-Bautista et al. (2017) [54]	W	Dynamic	AM	Kinect
Martínez-Guevara and Curiel (2024) [60]	W, P	Dynamic	AM	Camera/OpenPose
Martínez-Guevara et al. (2019) [61]	W	Dynamic	HC, AM	Camera
Trujillo-Romero and García-Bautista (2023) [55]	W, P	Dynamic	HC, AM	Kinect
Martínez-Guevara et al. (2023) [62]	W, P	Dynamic	HC, AM	Camera
Martínez-Sánchez et al. (2023) [20]	W	Dynamic	HC, AM, NHG	Camera
González-Rodríguez et al. (2024) [63]	P	Dynamic	HC, AM, NHG	Camera/MediaPipe
Miah et al. (2024) [35]	L, W, P	Dynamic	HC, AM, NHG	Camera/MediaPipe, OpenPose
Current study	W, P	Dynamic	AM	Camera/YOLOv8

* L: alphabet letter; N: number; W: word; P: phrase.

1.1. Toward a Recognition System for LSM

We present the sign data acquisition, the hardware selected, and the fundamental concepts of our research toward a recognition system for LSM.

1.1.1. Contactless Sign Data Acquisition with Simple Hardware

Due to the socioeconomic conditions of the main users of LSM, this research uses contactless, simple hardware for sign data acquisition (i.e., a pure vision-based approach), since color cameras are widely accessible and available in portable devices, which are very common in Mexico. As presented in Table 4, one important remark is that only one LSM research work [59] used contact sensing for sign data acquisition.

1.1.2. Sign Features

From a linguistics perspective, LSM signs present six documented parameters, that is, basic articulatory parameters that simultaneously combine to form signs [39,64–66]. We

propose a simplified Kinematics perspective, as shown in Figure 1, which combines four of those parameters into arm movements (AMs):

1. Hand configuration (HC): The shape adopted by one or both hands. As seen in Tables 1 and 3, most research focuses on HC only. Hand segmentation [67] and hand pose detectors are very promising technologies for this feature. The number of HCs required to perform a sign is variable in LSM; some examples regarding the number of HCs required for a sign are as follows: number "1" (1 HC), number "9" (2 HCs), number "15" (2 hands, 1 HC), and "grandmother" (2 hands, 3 HCs). See Appendix A for samples of these signs.
2. Non-hand gestures (NHGs): Facial expressions (frowning, raising eyebrows), gestures (puffing out cheeks, blowing), and body movements (pitching, nodding). While most signs do not require non-hand gestures, some LSM signs do. Some signs that require one or more NHGs are as follows: "How are you?", "I'm sorry", "Surprise!" (two NHGs of this sign are shown in Figure 1). See Appendix A for links to samples of these signs.
3. Arm movement (AM): This can be characterized by tracking the joint movements of wrists, shoulders, and elbows. It is enough to obtain the following basic articulatory parameters [39,64–66]:
    (a) Articulation location: This is the location on the signer's body or space where the signs are executed.
    (b) Hand movement: The type of movement made by the joints from one point to another.
    (c) Direction of movement: The trajectory followed by the hand when making the sign.
    (d) Hand orientation: Orientation of the palm of one or both hands, with respect to the signer's body when making the manual configuration.

    This part can be studied from pose-based approaches (c.f. [31,32] with pose estimation using AlphaPose).

Other decompositions have been proposed to simplify sign analysis, such as in [62] (Figure 1), where an LSM sign is decomposed into fixed postures and movements. We consider the fact that this approach could lose important information, as transitions in hand postures are also important (as documented in the Hamburg Notation System (HamNoSys)) [68].

The use of pose estimators, particularly MediaPipe, enables the extraction of facial, hand, and body features; c.f. [17,23]. The use of pose estimators is quite frequent in SL research, however, there are still areas for improvement (c.f. [27] (Figure 8)), where a PhBFC was designed to improve MediaPipe hand pose estimation. Complementary approaches like bimodal frameworks [17] highlight the current limitations of these estimators.

We believe that focusing on a single element to describe LSM is inadequate, given its meaning and contribution to the sign. But covering everything at the same time is also very complex, as seen in most LSM research. Since most of the LSM work focuses on HC, this paper focuses on the AM part and reports the approach created to analyze visual patterns in arm joint movements. Our current work uses YOLOv8 [69,70] for pose estimation. While it is a 2D method, and MediaPipe is better for 3D, we discuss our decision in Appendix B.

The main contribution of this work involves the use of arm movement keypoints, particularly wrist positions, as a partial feature for sign language recognition. This is motivated by the observation in [23], where wrist location played a crucial role in distinguishing similar signs. For instance, the same hand configuration used at different vertical positions (e.g., near the head to indicate a headache, or near the stomach to indicate a stomachache)

conveys different meanings. By isolating and analyzing this spatial feature, we aim to better understand its discriminative power in sign recognition tasks.

This paper is structured as follows. Section 2 describes the data acquisition, the experimental design and setup, the stages of the proposed approach for SLR, and the evaluation process and metrics. Section 3 describes the results from the analysis of two case studies and presents a comparison of the proposed methodology against state-of-the-art works. The potential and the limitations of our approach are discussed in Section 4. The conclusions of this work are presented in Section 5.

2. Materials and Methods

This section describes the resources, tools, and procedures used in this study. First, the acquisition of a visual sign language dataset is presented, including a detailed description of its features. Next, the experimental design is introduced, indicating the experiments to be performed and their objectives. Then, the experimental setup involving the deep learning models and the computational resources employed is described. Afterward, the stages of our first-step sign language recognition system are explained. Lastly, a detailed explanation of motion shapes used in the experimentation and the evaluation metrics is provided.

2.1. Data Acquisition

In this research, a proprietary dataset was developed with the active participation of the deaf community and LSM experts, ensuring no restrictions on recognizing hand configurations, arm movements, and facial expressions. The creation of the dataset was reviewed and approved by the Bioethics Committee for Human Research at Cinvestav, and all participants provided written informed consent.

The dataset comprises 74 signs—73 performed by 17 subjects and 1 ("iron") performed by 16 subjects. In total, we have 1257 color videos (900 × 720@90 fps) for RGB data acquisition. We consider this dataset a visual sign signal dataset.

All signs show HCs and AM, and three of them have NHGs ("How?", "How are you?", "Why?"). There are four phrases in the dataset, as follows: "Good morning!" ("*¡Buenos días!*"), "Good afternoon!" ("*¡Buenas tardes!*"), "How are you?" ("*¿Cómo estás?*"), and "Why?" ("*¿Por qué?*"). The latter is a question word in English, but it is constructed with two words in Spanish and, in LSM, is represented by a sign composed of two signs with independent meanings. This information is summarized in Table 5.

Table 5. Custom dataset.

Feature	Description
Signs *	70 W, 4 P
Signers	17
Samples	73 signs with 17 samples, 1 sign with 16 samples
Sign features	HC, AM, NHG
Sign signal	Visual
Signal properties	900 × 720@90 fps
File format	MKV videos
Samples for training	10 samples
Samples for validation	2 samples
Samples for testing	5 samples

* W: words; P: phrase.

2.2. Experimental Design

Experiments were conducted on the custom dataset. The goal of these experiments was to classify dynamic LSM signs by detecting and tracking the wrist, elbow, and shoulder joints in order to characterize the AM. For this purpose, since sign production involves motion and changes in shape in space, we decided to use a pose-based approach to transform the visual sign signals into keypoint sign signals, and CNN for classification.

Two case studies are presented in this research. The first case only considers shoulders and wrists, as the wrists exhibit the predominant movement while the shoulders serve as base joints with minimal displacement. The second case includes the elbows, in addition to the shoulders and wrists, as the elbows also experience significant movement.

To carry out these analyses, three groups of signs were selected from the custom dataset. Each group was chosen based on specific characteristics. The first two subsets were selected based on signs with visually distinguishable motion patterns; in contrast, the third subset is composed of signs with variants to examine how this variability influences the classifier's performance. More detailed information about these subsets is provided in Section 2.5.

2.3. Experimental Setup

For the experimentation, a pose detector and a CNN classifier framework were required. To select a pose estimation framework, we conducted preliminary experiments to compare the commonly used MediaPipe (Google LLC, Mountain View, CA, USA) and the YOLOv8-pose (Ultralytics Inc., Frederick, MD, USA) detector. Based on this comparison, we chose YOLOv8-pose due to its superior performance. The details of this comparison, which support our decision, can be found in Appendix B.

As YOLOv8-pose was selected for pose estimation, we used YOLOv8-cls (Ultralytics Inc., Frederick, MD, USA) to analyze visual patterns of the arm joint movements. Using a single technology for multiple tasks offers several advantages. For example, a unified architecture reduces the need for format adaptation between different models, simplifies implementation, and streamlines the workflow. Also, it reduces the possible problems of training and running multiple models across different frameworks.

A micromamba (QuantStack, Saint-Maur-des-Fossés, France) environment was employed for the installation and implementation of the pose detection and image classification models used in this work. Table 6 provides a summary of the technical specifications of the components of the experimental setting.

Table 6. Computational resources.

Component	Version/Model
Operating system	Ubuntu 22.04.2 (Canonical Ltd., London, England)
Graphics card	Asus ROG STRIX GeForce RTX 2080 Ti O11G (ASUS Holdings Mexico S.A. de C.V., Mexico City, Mexico)
Computing API	CUDA 12.4 (NVIDIA Corporation, Santa Clara, CA, USA)
Programming language	Python 3.11.8 (Python Software Foundation, Beaverton, OR, USA)
Machine Learning library	PyTorch 2.2.2 (Linux Foundation, San Francisco, CA, USA)
Framework	YOLO 8.1.47 (Ultralytics Inc., Frederick, MD, USA)

2.4. Sign Language Recognition

This work represents a preliminary step toward a recognition system, which comprises a three-step process, as follows: (1) pose estimation, (2) shape generation, and (3) class prediction. A simplified diagram of this process is presented in Figure 2.

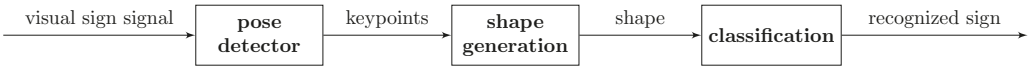


Figure 2. Pipeline of the arm movement approach for SLR.

In this setup, a video file is passed through a pose detector, where six keypoints are extracted for each frame and saved as a NumPy (NumFOCUS, Austin, TX, USA) array. These keypoints are then plotted to generate motion shapes, and the resulting images are used as input to a classification model. The classification model returns the top five predicted classes and their associated confidence score. Detailed descriptions of each stage in the process are provided in the following subsections.

2.4.1. Visual Sign Signals

To process the visual information, the video frames were cropped to 720 × 720 pixels (see Figure 3), as YOLOv8-pose operates internally on square images. This adjustment does not affect sign visibility, as all relevant joints remain within the square frame.



Figure 3. Dimensions of original and cropped frames.

2.4.2. Pose Estimation

LSM specifies that only the upper part of the body is meaningful in signing; so from the 17 keypoints detected by the selected pose detector, only 13 corresponding to the upper body are relevant; the 4 keypoints for knees and ankles are discarded. If the model fails to detect a joint, it is assigned a null value, which allows for easily discarding these missing values in further processing. Below is an example of pose estimation applied to the initial and final poses of the "deer" sign (Figure 4), as well as the extraction of the 13 keypoints.

The keypoints are stored in NPY format, a file type used by NumPy for efficiently storing data arrays. These arrays have dimensions of (13, 2, N): keypoints, 2D (x, y) coordinates, and the number of frames in each video.



Figure 4. Pose detection of the "deer" sign. (**Left**): neutral pose. (**Right**): final pose.

2.4.3. Shape Generation

From these arrays, the coordinates corresponding to the wrists, shoulders, and elbows are extracted according to each case study. The positions of these coordinates were plotted for each frame, illustrating the movement pattern of each joint, as shown in Figure 5.

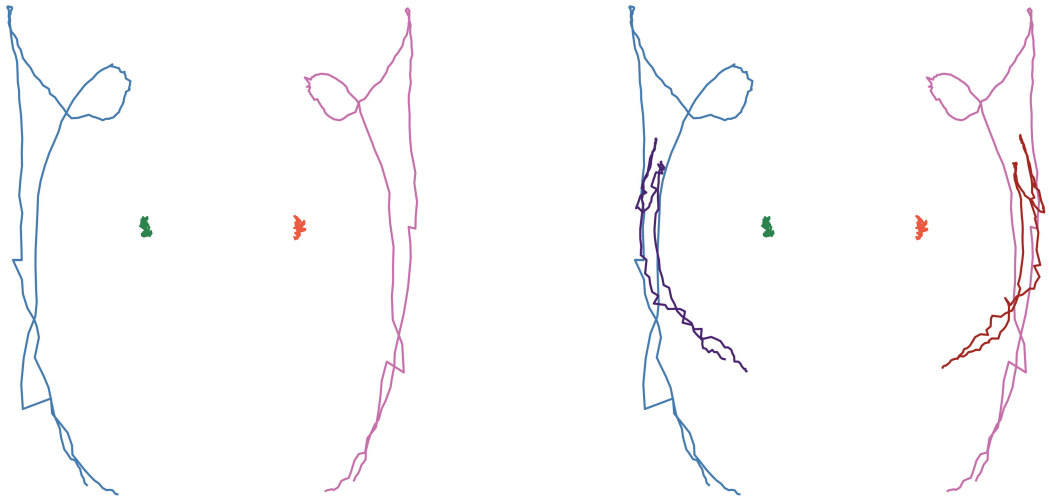


Figure 5. Movement shapes for the "deer" sign. (**Left**): only wrists and shoulder. (**Right**): also elbows.

2.4.4. Classification

The shape classification stage involves assigning each image a label from a predefined set of classes. For this purpose, the YOLOv8x-cls model was employed. This classifier is the most robust of the YOLOv8 classification models and maintains a deep CNN structure. The classifier outputs the top-5 predicted class labels along with their associated confidence scores.

The maximum number of examples per sign in all selected sets is 17; 10 examples were used for training, 2 for validation, and 5 for the testing phase. Table 7 shows the most relevant hyperparameters for model training and configuration. On the other hand, Table 8 details the data augmentation-related hyperparameters handled by YOLOv8 (not all parameters are active).

Table 7. Training parameters and their descriptions.

Parameter	Value	Description
epochs	50	Number of epochs or training cycles.
batch	16	Number of images processed in each iteration.
imgsz	224	Size of the images input into the model.
patience	100	Number of epochs without improvement before stopping the training.
lr0	0.01	Initial learning rate.
pre-trained	True	Indicates that the model uses pre-trained weights (ImageNet).
single_cls	False	If set to true, the model classifies into a single class.
dropout	0.0	Dropout rate. This is a regularization technique used to reduce overfitting in artificial neural networks.

Table 8. Image augmentation parameters and their descriptions.

Parameter	Value	Description
hsv_h	0.015	Hue of the image in the HSV color space.
hsv_s	0.7	Saturation of the image in the HSV color space.
hsv_v	0.4	Brightness of the image in the HSV color space.
degrees	0.0	Random rotation applied to the images.
translate	0.1	Random translation of the images.
scale	0.5	Random scaling factor applied to the images.
shear	0.0	Random shear angle applied to the images.
perspective	0.0	Perspective transformation applied to the images.
flipud	0.0	Probability of flipping the image vertically.
fliplr	0.5	Probability of flipping the image horizontally.
bgr	0.0	BGR to RGB color space correction factor.
mosaic	1.0	Probability of using the mosaic technique to combine images.
mixup	0.0	Probability of mixing two images.
copy_paste	0.0	Technique of copying and pasting objects between images.
auto_augment	randaugment	Type of data augmentation used.
erasing	0.4	Probability of erasing parts of the image to simulate occlusions.
crop_fraction	1.0	Proportion of the image to be cropped. A value of 1.0 indicates no cropping.

2.5. Evaluation

Experimentation was conducted on the two case studies outlined in Section 2.2, using three sets of motion shapes described below. The lists of signs in each subset are shown in Tables 9–11.

The first subset consists of a small group of five signs, chosen for their distinguishable shapes based on a qualitative evaluation. The primary objective of this group is to conduct a more controlled evaluation of the neural network, which allows for a clearer analysis of what the network is learning in an environment with fewer variables. Examples of these signs are presented in Figure 6, while the corresponding words are listed in Table 9.

Table 9. Signs for the first subset.

No.	Semantic Field	Sign
1	family	son *
2	greetings	hello *
3	days of the week	Monday *
4	family	godfather *
5	animals	deer *

* These signs are also in the second subset.

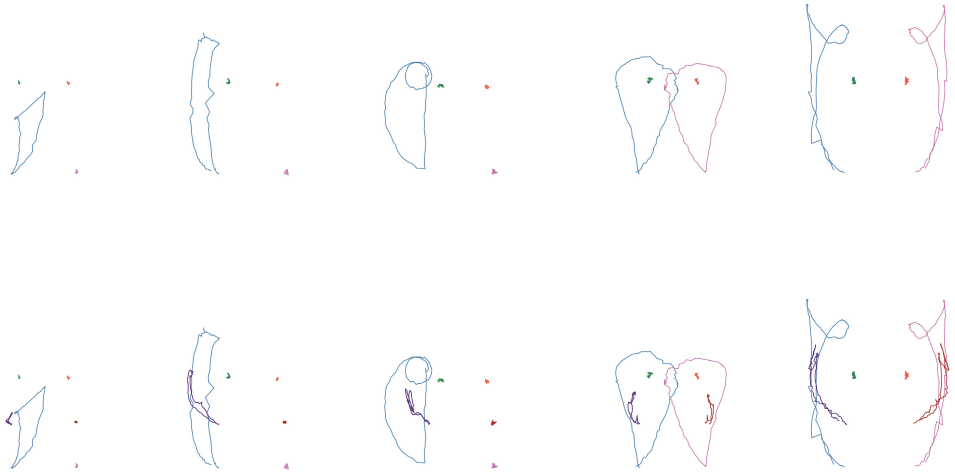


Figure 6. Shapes of the first subset (see words in Table 9). (**Top**): only wrists and shoulders. (**Bottom**): also elbows.

In the second subset, the signs are similarly distinguishable, but with a larger set consisting of 62 signs. The goal now is to assess whether the neural network's behavior remains consistent with that of the first set, despite the increased number of classes. Some examples of these signs are presented in Figure 7, and the corresponding words are listed in Table 10.

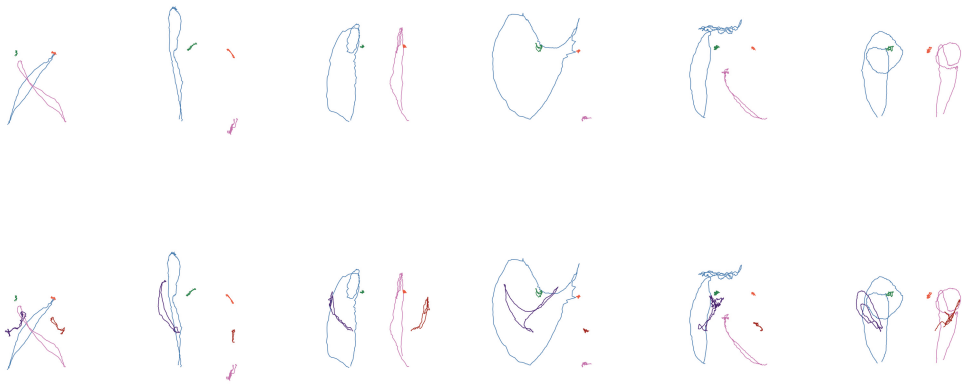


Figure 7. Shape examples of the second subset ("hug", "tall", "atole", "airplane", "flag", and "bicycle"). (**Top**): only wrists and shoulders. (**Bottom**): also elbows.

The third subset consists of 16 words related to the semantic field of house. This group is particularly notable for the high number of variants in its signs. As such, this experiment aims to assess the model's accuracy, as well as its ability to generalize and identify distinctive features within more complex sign language contexts. Examples of the sign forms from this set can be seen in Figure 8, and the corresponding vocabulary is outlined in Table 11.

Once the training stage is completed, the corresponding weights are saved in a custom model, which is then utilized for the subsequent testing phase. During this phase, key performance metrics, such as top-1 and top-5 accuracies, are collected. Top-1 accuracy measures how often the model's first prediction is correct, while top-5 accuracy evaluates whether the correct class appears among the five most probable predictions.

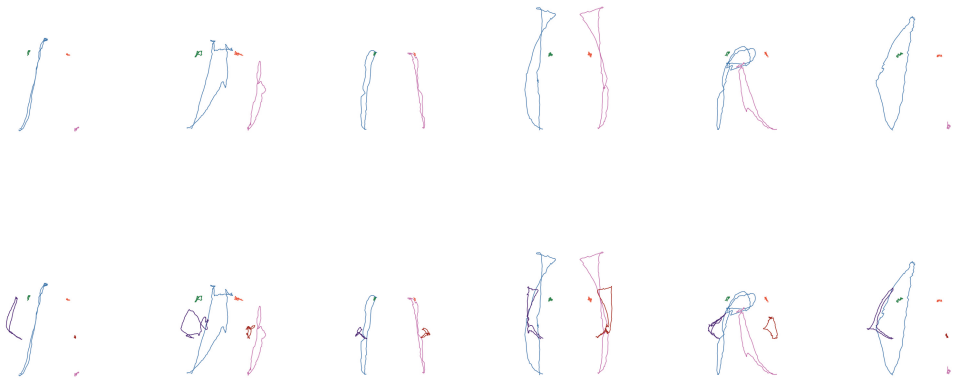


Figure 8. Shape examples of the third subset ("garbage", "trash can", "house", "curtains", "electricity", and "stairs"). (**Top**): only wrists and shoulders. (**Bottom**): also elbows.

Table 10. Signs for the second subset.

No.	Semantic Field	Sign	No.	Semantic Field	Sign
1	verbs	hug	32	verbs	to arrive
2	adjectives	tall	33	days of the week	Monday *
3	drinks	atole	34	kitchen	tablecloth
4	transport	airplane	35	miscellaneous	sea
5	school	flag	36	fruits	melon
6	transport	bicycle	37	kitchen	table
7	greetings	Good afternoon!	38	verbs	to swim
8	greetings	Good morning!	39	colors	dark
9	cities	capital	40	family	godfather *
10	house †	house	41	animals	bird
11	miscellaneous	sky	42	clothing	pants
12	questions	How?	43	animals	penguin
13	questions	How are you?	44	school	blackboard
14	school	classmate	45	food	pizza
15	house	curtains †	46	room	iron
16	days of the week	day	47	miscellaneous	please
17	house	broom †	48	questions	Why?
18	living room	light bulb	49	time	present
19	animals	rooster	50	professions	president
20	adjectives	fat	51	bathroom	shower
21	adjectives	big	52	living room	living room
22	verbs	to like	53	food	sauce
23	family	daughter	54	cities	Saltillo
24	family	son *	55	clothing	shorts
25	greetings	hello *	56	verbs	to dream
26	time	hour	57	transport	taxi
27	time	today	58	bathroom	towel
28	animals	giraffe	59	animals	deer *
29	verbs	to play	60	house	window †
30	drinks	milk	61	clothing	dress
31	vegetables	lettuce	62	person	widower

* These signs are also in the first training set. † These signs are also in the third subset.

Top-1 accuracy is computed using the standard approach employed for most classification tasks. It is defined as the proportion of correctly predicted labels over the total number of samples. Let $\hat{y}_i$ be the predicted label for the i-th sample and y_i the corresponding true label. The top-1 accuracy is then calculated as follows [71]:

$$\text{top-1}_{\text{accuracy}}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} \mathbf{1}(\hat{y}_i = y_i) \quad (1)$$

where $\mathbf{1}(x)$ is the indicator function, which returns 1 if the prediction is correct and 0 otherwise.

Table 11. Signs for the third subset.

No.	Semantic Field	Sign
1	house	garbage
2	house	trash can
3	house	house *
4	house	curtains *
5	house	electricity
6	house	stairs
7	house	broom *
8	house	internet
9	house	garden
10	house	keys
11	house	wall
12	house	floor
13	house	door
14	house	roof
15	house	mop
16	house	window *

* These signs are also in the second subset.

On the other hand, the top-k accuracy considers a prediction correct if the true label is among the k-highest predicted scores. Thus, top-1 accuracy is a special case of top-k accuracy, where $k = 1$.

Let $\hat{f}_{i,j}$ represent the predicted class for the i-th sample that has the j-th highest predicted score, and let y_i be the corresponding true label. The top-k accuracy is then calculated as follows [72]:

$$\text{top-k}_{\text{accuracy}}(y, \hat{f}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} \sum_{j=1}^{k} \mathbf{1}(\hat{f}_{i,j} = y_i) \quad (2)$$

where k is the number of top predictions considered, and $\mathbf{1}(x)$ is the indicator function.

These metrics are crucial for assessing the model's performance in a multi-class classification environment.

Additionally, a confusion matrix is generated for each experiment, providing a detailed overview of correct and incorrect predictions for each class. The results, along with their interpretation and analysis, are discussed in the following section.

3. Results

A total of seven SLR experiments on LSM were conducted (six with our custom dataset and one using an external dataset) to test our approach. The results are presented below.

Performance was evaluated using top-1 accuracy, top-5 accuracy, and the confusion matrix (see Section 2.5), which together provide a comprehensive view of the model's performance across each subset. In addition, performance graphs depicting loss and accuracy across training epochs are included, allowing observation of the model's learning curve over time.

3.1. Visual Sign Signal Dataset

3.1.1. First Subset

In the first experiment, five of the most distinguishable classes were selected (see confusion matrices in Figure 9). The results reveal that using only the shoulder and wrist coordinates achieved a top-1 accuracy of 0.9599. However, when the elbow coordinates were included, the top-1 accuracy decreased to 0.8799, suggesting that the additional information had a negative impact on performance.

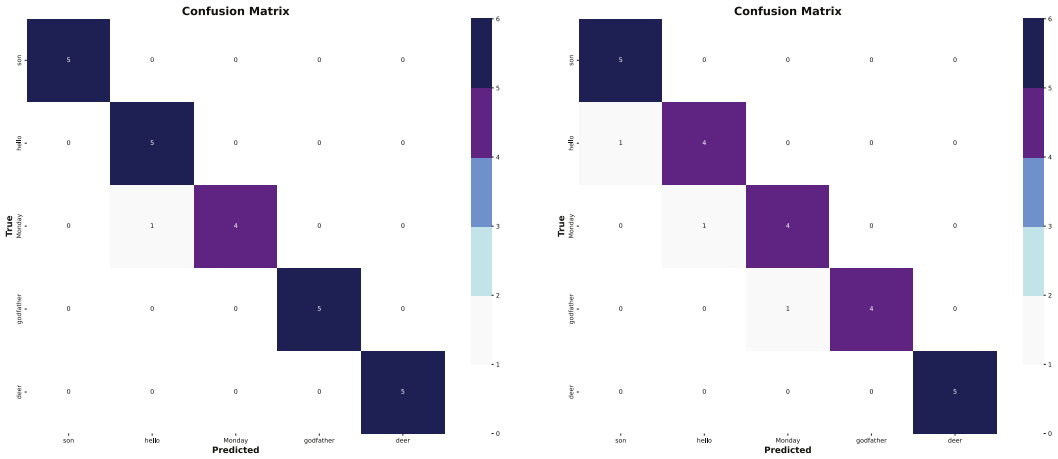


Figure 9. Confusion matrices for the first subset. (**Left**): only wrists and shoulders. (**Right**): also elbows.

Both the "son" and "deer" classes were classified with high accuracy in both case studies. However, slight confusion was observed between the "Monday" and "hello" classes in the first case. Additionally, when elbow coordinates were included, the model made errors in three of the five classes, indicating greater difficulty in differentiating between them. The performance graphs show that the accuracy in both models tends to stabilize around the 30th epoch, while the loss continues to decrease. Despite this, the model using only the wrist and shoulder coordinates outperformed the version with elbow coordinates, achieving higher accuracy (see graphs in Figure 10). In summary, the results are highly favorable in the best-case scenario, with a classification rate exceeding 95%. This suggests that the model is capable of effectively distinguishing between a limited number of well-defined classes. However, it is preferable to restrict the analysis to wrist and shoulder data, as including elbow data appears to negatively impact performance.

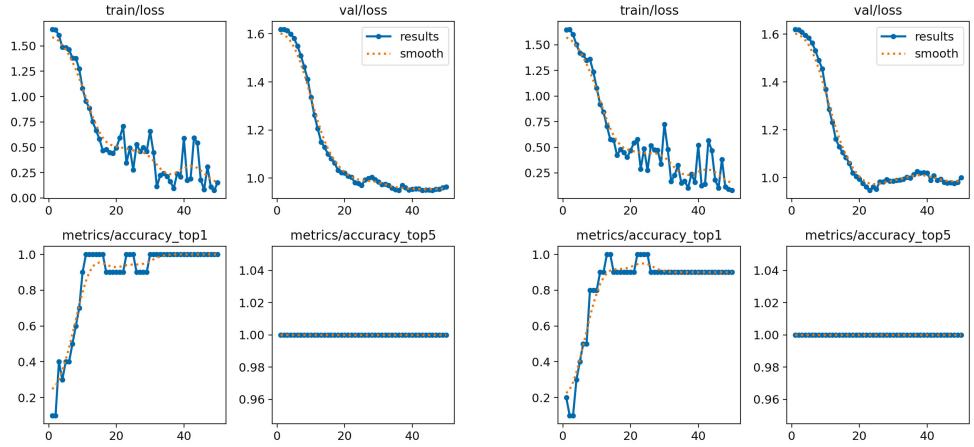


Figure 10. Performance charts for the first subset. (**Left**): only wrists and shoulders. (**Right**): also elbows.

3.1.2. Second Subset

In the second experiment, we expanded the number of classes to 62, while ensuring that the shapes remained distinguishable from one another (see confusion matrices in Figure 11). The model using only wrist and shoulder coordinates achieved a top-1 accuracy of 0.6375, whereas including elbow information resulted in a slight improvement to 0.6537.

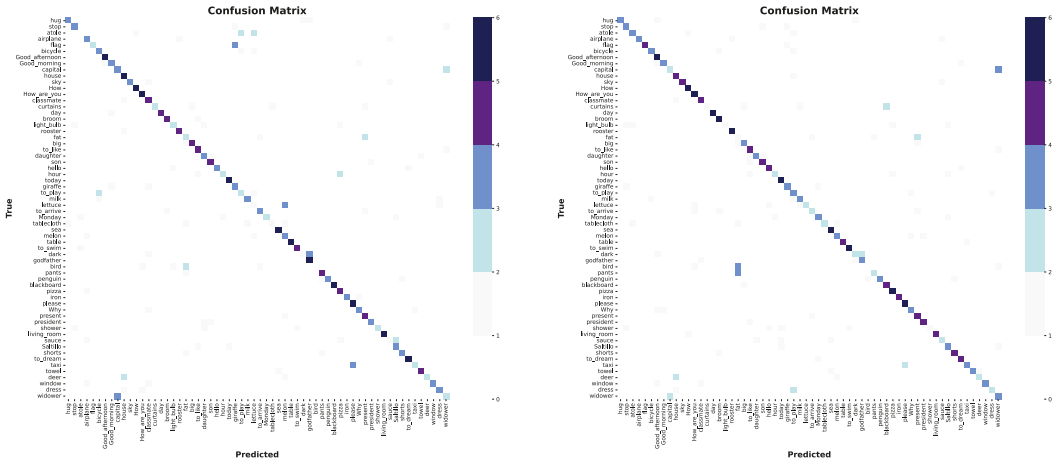


Figure 11. Confusion matrices for the second subset. (**Left**): only wrists and shoulders. (**Right**): also elbows.

For top-5 accuracy, the results were similar, with the first model achieving an accuracy of 0.8640, which improved to 0.8932 when elbow data was included. Performance analysis during training and validation revealed a consistent trend in both models, that is, accuracy steadily increased while loss progressively decreased (see Figure 12), indicating effective learning. The best model achieved an overall accuracy of 65%, which is acceptable, but showed variability in class performance. Some classes were classified nearly perfectly, while others exhibited notable precision issues. This suggests that, despite clear visual distinctions between classes, the large number of classes (62) combined with the limited number of examples per class (5) may hinder the model's ability to generalize effectively. In conclusion, although incorporating elbow information improves classification accuracy, the inconsistent performance underscores the need for more examples per class to optimize the model's results.

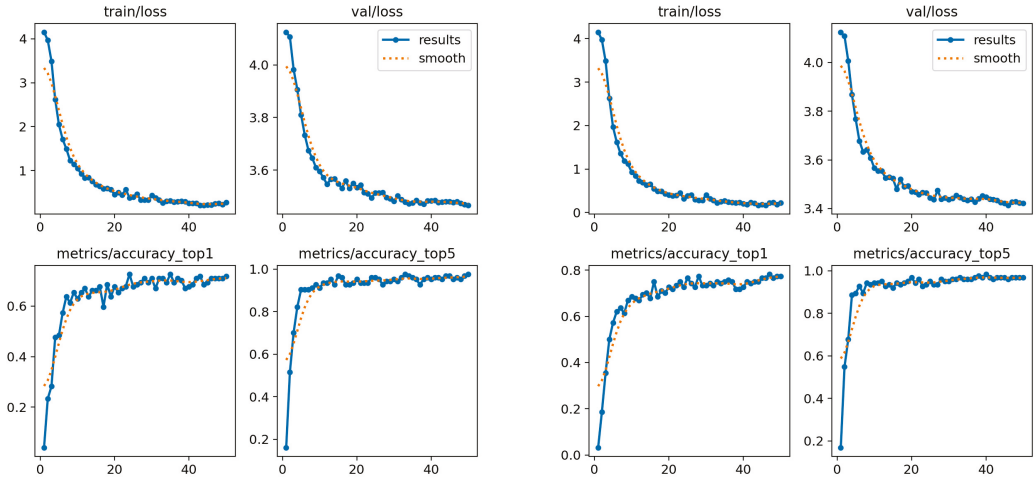


Figure 12. Performance charts for the second subset. (**Left**): only wrists and shoulders. (**Right**): also elbows.

3.1.3. Third Subset

In this experiment, the set is composed of 16 words in the home semantic field. The complexity of this group lies in the fact that some signs have variants. It is interesting to note that—in both models—words such as "internet", "keys", "mop", and "window" were classified correctly since they showed less variability. In contrast, words like "curtains", "garden", and "wall" performed poorly, with poor predictions in both models (see confusion matrices in Figure 13).

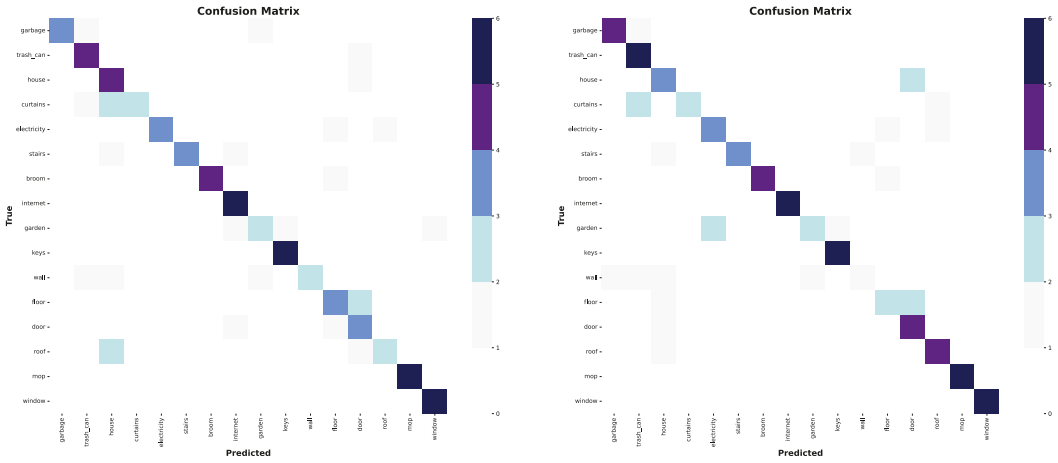


Figure 13. Confusion matrices for the third subset. (**Left**): only wrists and shoulders. (**Right**): also elbows.

The model using only wrist and shoulder information achieved a top-1 accuracy of 0.6875, while including the elbow coordinates increased the accuracy to 0.7125. For top-5 accuracy, both models achieved a value of 0.9250.

Performance in both studies was quite similar (see the graphs in Figure 14), showing fluctuations during training, but with a tendency to stabilize at a constant value toward the later stages. This suggests that the model managed to learn the main features of the characters, although its generalization capacity is limited by the complexity of the variants within the set. The classification rate reached up to 71% when the elbow information was included, which indicates that this additional information contributes positively to the recognition, although the increase in accuracy is not very significant.

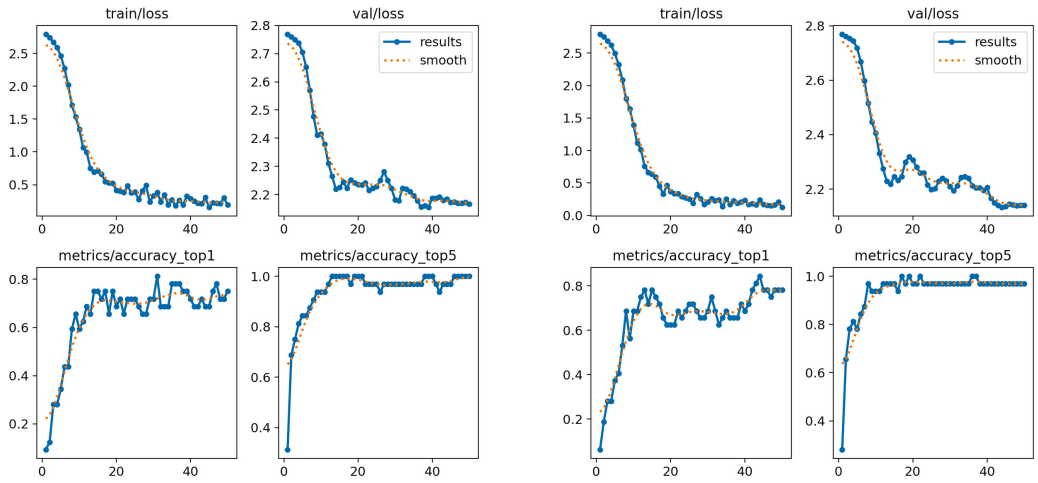


Figure 14. Performance charts for the third subset. (**Left**): only wrists and shoulders. (**Right**): also elbows.

Despite the limitations, the model was able to detect patterns in some cases. However, its ability to generalize across a large number of classes, variants, and a limited number

of examples is insufficient. Notwithstanding, the performance graphs reveal a tendency toward stabilization, suggesting that while the model holds potential for certain datasets, it requires additional information—such as finger movements—to enhance its classification accuracy in more complex scenarios.

3.2. Comparison of the Proposed Model on a Keypoint Sign Signal Dataset

In order to compare our approach against other state-of-the-art works, we needed to perform additional experiments on another LSM dataset. We selected from among the publicly available LSM datasets; see Tables 2 and 3. Our selection criterion was based on the number of SLR studies that used each dataset and reported performance accuracy, to enable a proper comparison. Therefore, we opted for a keypoint sign signal dataset, that is, the MSLR dataset from [23]. Details about this dataset are available in Appendix C. This dataset has been tested and reported by at least three different machine learning models [23,35,73]. In contrast to the visual sign signal dataset, the MSLR dataset required a shorter pipeline compared to the one required for visual signals. The pipeline is shown in Figure 15.

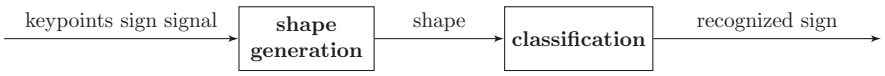


Figure 15. Pipeline of the arm movement approach for SLR with the MSLR dataset.

For this comparison, we tested our arm movement approach with this dataset using all the arm joints. For classification, we trained a model from scratch, using the current YOLO nanoarchitecture [74] YOLO11n-cls, with the PyTorch framework. This architecture uses 86 layers and has a computational complexity of 0.5 GFLOPs, with 1.633584 million parameters, when using a frame size of 224 pixels [75]. The results of this comparison are presented in Table 12.

Table 12. Performance accuracy with the MSLR dataset and a state-of-the-art comparison.

Ref.	Dataset	Joint Keypoints *	Performance Accuracy (%)
RNN [23]	MSLR	67	96.44
GRU [23]	MSLR	67	97.11
Dynamic-GCN † [73]	MSLR	67	98.55
Single-stream GCAR [35]	MSLR	67	99.00
Two-stream GCAR [35]	MSLR	67	99.75
Proposed model	MSLR	6	85.78

* 67 keypoints of the full body; 6 keypoints of the arm joints: wrists, elbows, and shoulders. † The model is presented in [73]; the performance accuracy is reported in [35].

4. Discussion

Table 13 presents the accuracy values based on the top-1 accuracy metric obtained using the YOLOv8x-cls model. The results indicate that including elbow coordinates led to better performance in two out of the three experiments. Although the improvement was modest (ranging from 3% to 4%), it suggests that incorporating additional joint information can contribute to more accurate classifications.

The experiments with various datasets allowed us to observe the behavior of the convolutional neural network (CNN) based on the input data. It became evident that the network's performance is heavily influenced by the selection of classes. Using all available classes from the database is not always ideal, as this tends to yield suboptimal

results. Therefore, a more focused approach, where only relevant classes are included, is recommended for improving model classification.

Table 13. Top-1 accuracy comparison on the custom dataset.

Dataset	No. Classes	Description	With Elbows	Without Elbows
1	5	More distinguishable	0.8799	0.9599
2	62	More or less distinguishable	0.6537	0.6375
3	16	House group	0.7125	0.6875

Despite certain limitations—such as the small number of examples per class, the presence of variants, and the high similarity between some signs—the neural network was still able to classify a significant number of signs correctly and recognize patterns in the movement data. This demonstrates the potential of the YOLOv8 model for this type of task.

Compared to other CNNs, YOLOv8 stands out due to its optimized architecture, which allows for the use of pre-trained models on large datasets like ImageNet. This enables the model to achieve high accuracy and efficiency, making it suitable for real-time applications. However, as with any model, performance is largely dependent on the quality and quantity of the input data. In this case, the limited number of examples (17 per class) restricts the network's ability to achieve optimal accuracy.

These results highlight both the potential and the limitations of our approach. The experiments demonstrated that it is possible to classify a considerable number of signs, indicating that this dataset and strategy could serve as a useful tool for training a convolutional neural network (CNN), such as YOLOv8. However, the analysis also reveals that the current structure of the dataset—characterized by a limited number of examples, variants between classes, and high similarity among some signs—presents challenges that must be addressed through alternative approaches.

The comparison between the two case studies (with and without elbows) was intended to assess whether the inclusion of a greater number of keypoints improves the performance of the model. This seems to indicate that this assumption is correct. The next immediate step is to optimize these results, either by using a different convolutional neural network (CNN) or by exploring different architectures, such as recurrent neural networks (RNNs), but keeping the focus on the use of keypoints; i.e., using pose-based approaches.

Additionally, the study performed on the MSLR dataset showed good results for the proposed approach (with an accuracy of 85.78 % using 6 keypoints), compared to the extraordinary results obtained by [23] (with accuracies of 96.44 % and 97.11 %) and [35] (with accuracies of 99 % and 99.75 %) while using the complete keypoint sign signals in the dataset; see Table 12. This is an interesting finding that shows the relevance of AM sign features, as most previous research studies typically focused on HC sign features.

5. Conclusions

This paper presents ongoing work toward the creation of a recognition system for LSM. A decomposition of sign features is proposed into HC, AM, and NHG. Contactless, simple hardware was used for sign signal acquisition. A custom proprietary dataset of 74 signs (70 words and 4 phrases) was constructed for this research. In contrast to most LSM research, this paper reports an analysis focused on the AM part of signs, rather than on HC-focused or holistic approaches (HC + AM + NHG).

The analysis was conducted through a series of classification experiments using YOLOv8, aimed at identifying visual patterns in the movement of key joints, i.e., wrists, shoulders, and elbows. A pose detection model was used to extract joint movements,

followed by an image classification model (both integrated into YOLOv8) to classify the shapes generated by these movements.

These experiments are the first stage of a larger project. For now, we are focusing on the analysis of arm movement (shoulders, elbows, and wrists) because it is a less-studied feature, and information can be extracted from it using a relatively simple methodology.

Later, the goal will be to integrate other essential components of sign language, such as manual configuration and non-hand gestures, to develop a more complete system. Ultimately, this will support progress toward automatic sign language recognition.

Author Contributions: Conceptualization, G.H.-A., K.O.-H. and M.C.; methodology, G.H.-A., K.O.-H. and M.C.; software, G.H.-A. and K.O.-H.; validation, G.H.-A.; formal analysis, G.H.-A., K.O.-H. and M.C.; investigation, G.H.-A., K.O.-H. and M.C.; resources, K.O.-H. and M.C.; data curation, G.H.-A.; writing—original draft preparation, G.H.-A., K.O.-H. and I.L.-J.; writing—review and editing, G.H.-A., K.O.-H., M.C. and I.L.-J.; visualization, G.H.-A.; supervision, K.O.-H. and M.C.; project administration, K.O.-H.; funding acquisition, G.H.-A. and I.L.-J. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by CONAHCYT through scholarship grant number 828990.

Institutional Review Board Statement: Ethical review and approval were conducted by the Ethics Committee of Cinvestav (protocol code: 105/2023; date of approval: 7 December 2023).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The datasets presented in this article are not readily available because the data are part of an ongoing study, and due to technical and time limitations. Requests to access the datasets should be directed to keny.ordaz@cinvestav.edu.mx.

Acknowledgments: We thank Felipe Hernández Rodríguez for providing a space at his institution for dataset acquisition. We thank Hilda Xóchitl Cabrera Hernández; Daniela Fernanda Espinoza Ibarra; and María Guadalupe Luna Arguello for their help with contacting participants.

Conflicts of Interest: The authors declare no conflicts of interest. The funder had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

AM	arm movement
API	application programming interface
ArSL	Arabic Sign Language
ASL	American Sign Language
BdSL	Bangladeshi Sign Language
CESAR	Recife Center for Advanced Studies and Systems
CSL	Chinese Sign Language
CSV	comma-separated values
CNN	convolutional neural network
DGS	German Sign Language (*Deutsche Gebärdensprache*)
EMG	electromyography
FLV	flash video
fps	frames per second, frame rate
GCAR	graph convolution with attention and residual connection
GCN	graph convolutional network
GRU	gated recurrent units

HC	hand configuration
IMU	inertial measurement unit
JPEG	Joint Photographic Experts Group, ISO/IEC 10918
LIBRAS	Brazilian Sign Language (*Língua Brasileira de Sinais*)
LSA	Argentinian Sign Language (*Lengua de Señas Argentina*)
LSE	Spanish Sign Language (*Lengua de Señas Española*)
LSM	Mexican Sign Language (*Lengua de Señas Mexicana*)
LSTM	long short-term memory
MKV	Matroska video
MLP	multilayer perceptron
MSLR	Mexican Sign Language Recognition dataset
MP4	MPEG-4 Part 14, ISO/IEC 14496-14:2003
NHG	non-hand gesture
NN	neural network
NPY	NumPy standard binary file format
PJM	Polish Sign Language (*Polski Język Migowy*)
PSL	Pakistan Sign Language
RGBD	red, green, blue, and depth
RNN	recurrent neural network
sEMG	surface EMG
SL	sign language
SLR	sign language recognition
SVM	support vector machine
SWF	small web format
TSL	Turkish Sign Language
YOLO	you only look once

Appendix A. Digital Glossary of LSM

The GDLSM [38] has 747 signs grouped into 19 thematic categories. We provide direct links to some of the signs included in this digital glossary, which were mentioned in Section 1.1.2.

- Numbers (*Números*); https://lsm.indiscapacidad.cdmx.gob.mx/ejes/numeros/, accessed on 5 June 2025:
  - 1 (one): https://youtu.be/zcd4GfYz-fA, accessed on 5 June 2025.
  - 9 (nine): https://youtu.be/MgnvumQM-cQ, accessed on 5 June 2025.
  - 15 (fifteen) (first variant): https://youtu.be/yZ3X38cFWUM, accessed on 5 June 2025.
  - 15 (fifteen) (second variant): https://youtu.be/64jBCZXv6rY, accessed on 5 June 2025.
- Family (*Familia*); https://lsm.indiscapacidad.cdmx.gob.mx/ejes/familia/, accessed on 5 June 2025:
  - Grandmother (*abuela*) : https://youtu.be/lckOvtr0lZU, accessed on 5 June 2025.
- Everyday expressions (*Expresiones cotidianas*); https://lsm.indiscapacidad.cdmx.gob.mx/ejes/expresiones-cotidianas/, accessed on 5 June 2025:
  - How are you? (*¿Cómo estás?*) https://youtu.be/x7zFMacTe04, accessed on 5 June 2025.
  - I'm sorry (*Disculpa*) https://youtu.be/bWwIisAtYCI, accessed on 5 June 2025.
  - Surprise! (*¡Sorpresa!*) https://youtu.be/Q0OqTBjoIjU, accessed on 5 June 2025, this sign was used in Figure 1.

Appendix B. Comparison Between MediaPipe and YOLOv8 Pose Detection Models

MediaPipe detects 33 keypoints with its Pose Landmarker (Heavy) model, and it can provide 2D and 3D coordinates. YOLOv8 detects 17 keypoints with its YOLOv8x-pose-p6 model and provides 2D coordinates. YOLOv8 keypoints 5–10 are for the shoulder, elbow, and wrist joints, and MediaPipe keypoints 11–16 are for the same joints. We compared the MediaPipe and the YOLOv8 pose detectors in several signs. We decided to use YOLOv8 over MediaPipe due to frequent tracking failures of the wrist joint in many of the signs, particularly in the occluded conditions of the hands. An example of this issue is shown in Figure A1.

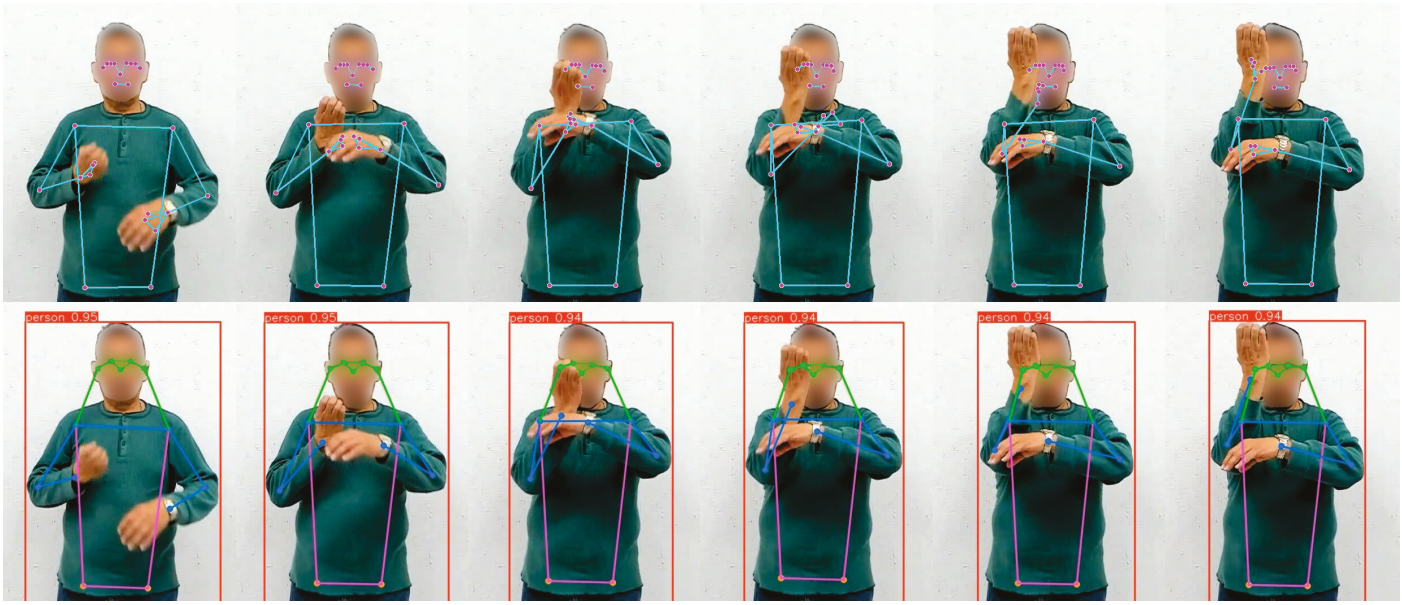


Figure A1. Comparison of wrist joint tracking between YOLOv8 and MediaPipe. Example with the "state" sign. Top row: MediaPipe. Bottom row: YOLOv8 pose detector. Four inner frames: MediaPipe loses track of the wrist joint, while YOLOv8 keeps track of the AM in all frames.

Appendix C. MSLR Dataset

The Mexican Sign Language Recognition (MSLR) dataset was created by [23]. It contains samples of 30 signs in LSM; see Table A1. This is a keypoint-based sign signal dataset, as shown in Table A2. Each sample consists of 20 frames, with 67 pose keypoints recorded per frame. The keypoints are distributed as follows: 20 for the face, 5 for the body (shoulders, elbows, and a midpoint between the shoulders), and 21 for each hand.

Table A1. Signs in the MSLR [23] dataset.

No.	Group *	Sign
1	alphabet	A
2	alphabet	B
3	alphabet	C
4	alphabet	D
5	alphabet	J
6	alphabet	K
7	alphabet	Q
8	alphabet	X
9	questions	What?
10	questions	When?

Table A1. *Cont.*

No.	Group *	Sign
11	questions	How much?
12	questions	Where?
13	questions	For what?
14	questions	Why?
15	questions	What is that?
16	questions	Who?
17	days of the week	Monday
18	days of the week	Tuesday
19	days of the week	Wednesday
20	days of the week	Thursday
21	days of the week	Friday
22	days of the week	Saturday
23	days of the week	Sunday
24	frequent words	(to) spell
25	frequent words	(to) explain
26	frequent words	thank you
27	frequent words	name
28	frequent words	please
29	frequent words	yes
30	frequent words	no

* Group names and information taken from [23] (Table 2).

Table A2. MSLR [23] dataset.

Feature	Description
Signs *	8 L, 21 W, 1 P
Signers	4
Samples	30 signs with 100 samples
Sign features	HC, AM, NHG
Sign signal	Keypoints
File format	CSV files
Samples for training †	70 samples
Samples for validation †	15 samples
Samples for testing †	15 samples

* L: letters, W: words; P: phrase. † This split was defined by the dataset authors.

References

1. World Health Organization. World Reporting on Hearing. 2021. Available online: https://www.who.int/publications/i/item/9789240020481 (accessed on 31 March 2025).
2. Secretaría de Salud. 530. Con Discapacidad Auditiva, 2.3 Millones de Personas: Instituto Nacional de Rehabilitación. 2021. Available online: https://www.gob.mx/salud/prensa/530-con-discapacidad-auditiva-2-3-millones-de-personas-instituto-nacional-de-rehabilitacion (accessed on 31 March 2025).
3. SLAIT. SLAIT—AI-Driven American Sign Language Translator. 2024. Available online: https://slait.ai (accessed on 29 March 2025).
4. Lenovo. Lenovo's AI-Powered Sign Language Translation Solution Empowers Signers in Brazil. 2023. Available online: https://news.lenovo.com/ai-powered-sign-language-translation-solution-hearing-brazil/ (accessed on 31 March 2025).
5. Rocha, J.V.; Lensk, J.; Ferreira, M.D.C. Techniques for Determining Sign Language Gesture Partially Shown in Image(s). United States Patent 11587362B2, 21 February 2023.
6. Mane, V.; Puniwala, S.N.; Rane, V.N.; Gurav, P. Advancements in Sign Language Recognition: Empowering Communication for Individuals with Speech Impairments. *Grenze Int. J. Eng. Technol. (GIJET)* **2024**, *10*, 4978–4984.
7. Krishnan, S.R.; Varghese, C.M.; Jayaraj, A.; Nair, A.S.; Joshy, D.; Sulbi, I.N. Advancements in Sign Language Recognition: Dataset Influence on Model Accuracy. In Proceedings of the 2024 International Conference on IoT Based Control Networks and Intelligent Systems (ICICNIS), Bengaluru, India, 17–18 December 2024; pp. 1563–1568. [CrossRef]

8. Chiradeja, P.; Liang, Y.; Jettanasen, C. Sign Language Sentence Recognition Using Hybrid Graph Embedding and Adaptive Convolutional Networks. *Appl. Sci.* **2025**, *15*, 2957. [CrossRef]
9. Umut, I.; Kumdereli, U.C. Novel Wearable System to Recognize Sign Language in Real Time. *Sensors* **2024**, *24*, 4613. [CrossRef] [PubMed]
10. Rodríguez-Tapia, B.; Ochoa-Zezzatti, A.; Marrufo, A.I.S.; Arballo, N.C.; Carlos, P.A. Sign Language Recognition Based on EMG Signals through a Hibrid Intelligent System. *Res. Comput. Sci.* **2019**, *148*, 253–262. [CrossRef]
11. Gu, Y.; Oku, H.; Todoh, M. American Sign Language Recognition and Translation Using Perception Neuron Wearable Inertial Motion Capture System. *Sensors* **2024**, *24*, 453. [CrossRef]
12. Filipowska, A.; Filipowski, W.; Mieszczanin, J.; Bryzik, K.; Henkel, M.; Skwarek, E.; Raif, P.; Sieciński, S.; Doniec, R.; Mika, B.; et al. Pattern Recognition in the Processing of Electromyographic Signals for Selected Expressions of Polish Sign Language. *Sensors* **2024**, *24*, 6710. [CrossRef]
13. Galván-Ruiz, J.; Travieso-González, C.M.; Pinan-Roescher, A.; Alonso-Hernández, J.B. Robust Identification System for Spanish Sign Language Based on Three-Dimensional Frame Information. *Sensors* **2023**, *23*, 481. [CrossRef]
14. Hao, Z.; Duan, Y.; Dang, X.; Liu, Y.; Zhang, D. Wi-SL: Contactless Fine-Grained Gesture Recognition Uses Channel State Information. *Sensors* **2020**, *20*, 4025. [CrossRef]
15. Wang, Y.; Hao, Z.; Dang, X.; Zhang, Z.; Li, M. UltrasonicGS: A Highly Robust Gesture and Sign Language Recognition Method Based on Ultrasonic Signals. *Sensors* **2023**, *23*, 1790. [CrossRef]
16. Al-Saidi, M.; Ballagi, A.; Hassen, O.A.; Saad, S.M. Type-2 Neutrosophic Markov Chain Model for Subject-Independent Sign Language Recognition: A New Uncertainty–Aware Soft Sensor Paradigm. *Sensors* **2024**, *24*, 7828. [CrossRef]
17. Gao, Q.; Hu, J.; Mai, H.; Ju, Z. Holistic-Based Cross-Attention Modal Fusion Network for Video Sign Language Recognition. *IEEE Trans. Comput. Soc. Syst.* **2024**, *early access.* [CrossRef]
18. Lugaresi, C.; Tang, J.; Nash, H.; McClanahan, C.; Uboweja, E.; Hays, M.; Zhang, F.; Chang, C.L.; Yong, M.; Lee, J.; et al. MediaPipe: A Framework for Perceiving and Processing Reality. In Proceedings of the Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019, Long Beach, CA, USA, 17 June 2019.
19. Serafín De Fleischmann, M.; González Pérez, R. *Manos con voz: Diccionario de Lengua de Señas Mexicana*; Consejo Nacional para Prevenir la Discriminación: Mexico City, Mexico, 2011.
20. Martínez-Sánchez, V.; Villalón-Turrubiates, I.; Cervantes-Álvarez, F.; Hernández-Mejía, C. Exploring a Novel Mexican Sign Language Lexicon Video Dataset. *Multimodal Technol. Interact.* **2023**, *7*, 83. [CrossRef]
21. Espejel-Cabrera, J.; Dominguez, L.; Cervantes, J.; Cervantes, J. Mexican Sign Language Dataset. 2023. Available online: https://data.mendeley.com/datasets/6rj76z6y3n/1 (accessed on 31 March 2025). [CrossRef]
22. Espejel, J.; Jalili, L.D.; Cervantes, J.; Canales, J.C. Sign language images dataset from Mexican sign language. *Data Brief* **2024**, *55*, 110566. [CrossRef]
23. Mejía-Peréz, K.; Córdova-Esparza, D.M.; Terven, J.; Herrera-Navarro, A.M.; García-Ramírez, T.; Ramírez-Pedraza, A. Automatic Recognition of Mexican Sign Language Using a Depth Camera and Recurrent Neural Networks. *Appl. Sci.* **2022**, *12*, 5523. [CrossRef]
24. Yao, D.; Wang, W.; Wang, H.; Luo, Y.; Ding, H.; Gu, Y.; Wu, H.; Tao, K.; Yang, B.R.; Pan, S.; et al. Ultrasensitive and Breathable Hydrogel Fiber-Based Strain Sensors Enabled by Customized Crack Design for Wireless Sign Language Recognition. *Adv. Funct. Mater.* **2025**, *35*, 2416482. [CrossRef]
25. Wei, C.; Liu, S.; Yuan, J.; Zhu, R. Multimodal hand/finger movement sensing and fuzzy encoding for data-efficient universal sign language recognition. *InfoMat* **2025**, *7*, e12642. [CrossRef]
26. Wang, W.; Bo, X.; Li, W.; Eldaly, A.B.M.; Wang, L.; Li, W.J.; Chan, L.L.H.; Daoud, W.A. Triboelectric Bending Sensors for AI-Enabled Sign Language Recognition. *Adv. Sci.* **2025**, *12*, 2408384. [CrossRef]
27. Urrea, C.; Kern, J.; Navarrete, R. Bioinspired Photoreceptors with Neural Network for Recognition and Classification of Sign Language Gesture. *Sensors* **2023**, *23*, 9646. [CrossRef]
28. Niu, P. Convolutional neural network for gesture recognition human-computer interaction system design. *PLoS ONE* **2025**, *20*, e0311941. [CrossRef]
29. Raihan, M.J.; Labib, M.I.; Jim, A.A.J.; Tiang, J.J.; Biswas, U.; Nahid, A.A. Bengali-Sign: A Machine Learning-Based Bengali Sign Language Interpretation for Deaf and Non-Verbal People. *Sensors* **2024**, *24*, 5351. [CrossRef]
30. Woods, L.T.; Rana, Z.A. Modelling Sign Language with Encoder-Only Transformers and Human Pose Estimation Keypoint Data. *Mathematics* **2023**, *11*, 2129. [CrossRef]
31. Eunice, J.; J, A.; Sei, Y.; Hemanth, D.J. Sign2Pose: A Pose-Based Approach for Gloss Prediction Using a Transformer Model. *Sensors* **2023**, *23*, 2853. [CrossRef] [PubMed]
32. Kim, Y.; Baek, H. Preprocessing for Keypoint-Based Sign Language Translation without Glosses. *Sensors* **2023**, *23*, 3231. [CrossRef] [PubMed]

33. Boháček, M.; Hrúz, M. Sign Pose-based Transformer for Word-level Sign Language Recognition. In Proceedings of the 2022 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW), Waikoloa, HI, USA, 4–8 January 2022; pp. 182–191. [CrossRef]
34. Cihan Camgöz, N.; Koller, O.; Hadfield, S.; Bowden, R. Sign Language Transformers: Joint End-to-End Sign Language Recognition and Translation. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10020–10030. [CrossRef]
35. Miah, A.S.M.; Hasan, M.A.M.; Nishimura, S.; Shin, J. Sign Language Recognition Using Graph and General Deep Neural Network Based on Large Scale Dataset. *IEEE Access* **2024**, *12*, 34553–34569. [CrossRef]
36. Gil-Martín, M.; Villa-Monedero, M.; Pomirski, A.; Sáez-Trigueros, D.; San-Segundo, R. Sign Language Motion Generation from Sign Characteristics. *Sensors* **2023**, *23*, 9365. [CrossRef]
37. Villa-Monedero, M.; Gil-Martín, M.; Sáez-Trigueros, D.; Pomirski, A.; San-Segundo, R. Sign Language Dataset for Automatic Motion Generation. *J. Imaging* **2023**, *9*, 262. [CrossRef]
38. INDISCAPACIDAD. Glosario Digital de Lengua de Señas Mexicana. 2024. Available online: https://lsm.indiscapacidad.cdmx.gob.mx (accessed on 31 March 2025).
39. Calvo-Hernández, M.T. Diccionario Español-Lengua de Señas Mexicana (DIELSEME). 2004. Available online: http://campusdee.ddns.net/dielseme.aspx (accessed on 31 March 2025).
40. Álvarez Hidalgo, A.; Acosta-Arellano, A.; Moctezuma-Contreras, C.; Sanabria-Ramos, E. Diccionario Lengua de Señas Mexicana (DIELSEME 2). 2009. Available online: http://campusdee.ddns.net/dielseme.aspx (accessed on 31 March 2025).
41. Cruz-Aldrete, M. Hacia la construcción de un diccionario de Lengua de Señas Mexicana. *Rev. de Investig.* **2014**, *38*, 57–80.
42. Solís, F.; Martínez, D.; Espinoza, O. Automatic Mexican Sign Language Recognition Using Normalized Moments and Artificial Neural Networks. *Engineering* **2016**, *8*, 733–740. [CrossRef]
43. Carmona-Arroyo, G.; Rios-Figueroa, H.V.; Avendaño-Garrido, M.L. Mexican Sign-Language Static-Alphabet Recognition Using 3D Affine Invariants. In *Machine Vision Inspection Systems, Volume 2: Machine Learning-Based Approaches*; Scrivener Publishing LLC: Beverly, MA, USA, 2021; pp. 171–192. [CrossRef]
44. Salinas-Medina, A.; Neme-Castillo, J.A. A real-time deep learning system for the translation of mexican signal language into text. In Proceedings of the 2021 Mexican International Conference on Computer Science (ENC), Morelia, Mexico, 9–11 August 2021; pp. 1–7. [CrossRef]
45. Rios-Figueroa, H.V.; Sánchez-García, A.J.; Sosa-Jiménez, C.O.; Solís-González-Cosío, A.L. Use of Spherical and Cartesian Features for Learning and Recognition of the Static Mexican Sign Language Alphabet. *Mathematics* **2022**, *10*, 2904. [CrossRef]
46. Morfín-Chávez, R.F.; Gortarez-Pelayo, J.J.; Lopez-Nava, I.H. Fingerspelling Recognition in Mexican Sign Language (LSM) Using Machine Learning. In *Advances in Computational Intelligence*; Calvo, H.; Martínez-Villaseñor, L.; Ponce, H., Eds.; Springer: Cham, Switzerland, 2023; pp. 110–120. [CrossRef]
47. Sánchez-Vicinaiz, T.J.; Camacho-Pérez, E.; Castillo-Atoche, A.A.; Cruz-Fernandez, M.; García-Martínez, J.R.; Rodríguez-Reséndiz, J. MediaPipe Frame and Convolutional Neural Networks-Based Fingerspelling Detection in Mexican Sign Language. *Technologies* **2024**, *12*, 124. [CrossRef]
48. García-Gil, G.; López-Armas, G.d.C.; Sánchez-Escobar, J.J.; Salazar-Torres, B.A.; Rodríguez-Vázquez, A.N. Real-Time Machine Learning for Accurate Mexican Sign Language Identification: A Distal Phalanges Approach. *Technologies* **2024**, *12*, 152. [CrossRef]
49. Jimenez, J.; Martin, A.; Uc, V.; Espinosa, A. Mexican Sign Language Alphanumerical Gestures Recognition using 3D Haar-like Features. *IEEE Lat. Am. Trans.* **2017**, *15*, 2000–2005. [CrossRef]
50. Martinez-Seis, B.; Pichardo-Lagunas, O.; Rodriguez-Aguilar, E.; Saucedo-Diaz, E.R. Identification of Static and Dynamic Signs of the Mexican Sign Language Alphabet for Smartphones using Deep Learning and Image Processing. *Res. Comput. Sci.* **2019**, *148*, 199–211. [CrossRef]
51. Martínez-Gutiérrez, M.E.; Rojano-Cáceres, J.R.; Benítez-Guerrero, E.; Sánchez-Barrera, H.E. Data Acquisition Software for Sign Language Recognition. *Res. Comput. Sci.* **2019**, *148*, 205–211. [CrossRef]
52. Rodriguez, M.; Oubram, O.; Ali, B.; Lakouari, N. Mexican Sign Language's Dactylology and Ten First Numbers–Extracted Features and Models. 2023. Available online: https://data.mendeley.com/datasets/hmsc33hmkz/1 (accessed on 31 March 2025). [CrossRef]
53. Sosa-Jiménez, C.O.; Ríos-Figueroa, H.V.; Rechy-Ramírez, E.J.; Marin-Hernandez, A.; González-Cosío, A.L.S. Real-time Mexican Sign Language recognition. In Proceedings of the 2017 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC), Ixtapa, Mexico, 8–10 November 2017; pp. 1–6. [CrossRef]
54. García-Bautista, G.; Trujillo-Romero, F.; Caballero-Morales, S.O. Mexican Sign Language Recognition Using Kinect and Data Time Warping Algorithm. In Proceedings of the 2017 International Conference on Electronics, Communications and Computers (CONIELECOMP), Cholula, Mexico, 22–24 February 2017; pp. 1–5. [CrossRef]
55. Trujillo-Romero, F.; García-Bautista, G. Mexican Sign Language Corpus: Towards an Automatic Translator. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* **2023**, *22*, 212. [CrossRef]

56. Espejel-Cabrera, J.; Cervantes, J.; García-Lamont, F.; Ruiz-Castilla, J.S.; Jalili, Dominguez, L. Mexican sign language segmentation using color based neuronal networks to detect the individual skin color. *Expert Syst. Appl.* **2021**, *183*, 115295. [CrossRef]
57. Rodriguez, M.; Oubram, O.; Bassam, A.; Lakouari, N.; Tariq, R. Mexican Sign Language Recognition: Dataset Creation and Performance Evaluation Using MediaPipe and Machine Learning Techniques. *Electronics* **2025**, *14*, 1423. [CrossRef]
58. Sosa-Jiménez, C.O.; Ríos-Figueroa, H.V.; Solís-González-Cosío, A.L. A Prototype for Mexican Sign Language Recognition and Synthesis in Support of a Primary Care Physician. *IEEE Access* **2022**, *10*, 127620–127635. [CrossRef]
59. Varela-Santos, H.; Morales-Jiménez, A.; Córdova-Esparza, D.M.; Terven, J.; Mirelez-Delgado, F.D.; Orenday-Delgado, A. Assistive Device for the Translation from Mexican Sign Language to Verbal Language. *Comput. y Sist.* **2021**, *25*, 451–464. [CrossRef]
60. Martínez-Guevara, N.; Curiel, A. Quantitative Analysis of Hand Locations in both Sign Language and Non-linguistic Gesture Videos. In Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources, Turin, Italy, 20 May 2024; pp. 225–234.
61. Martínez-Guevara, N.; Rojano-Cáceres, J.R.; Curiel, A. Detection of Phonetic Units of the Mexican Sign Language. In Proceedings of the 2019 International Conference on Inclusive Technologies and Education (CONTIE), San Jose del Cabo, Mexico, 30 October–1 November 2019; pp. 168–1685. [CrossRef]
62. Martínez-Guevara, N.; Rojano-Cáceres, J.R.; Curiel, A. Unsupervised extraction of phonetic units in sign language videos for natural language processing. *Univers. Access Inf. Soc.* **2023**, *22*, 1143–1151. [CrossRef]
63. González-Rodríguez, J.R.; Córdova-Esparza, D.M.; Terven, J.; Romero-González, J.A. Towards a Bidirectional Mexican Sign Language–Spanish Translation System: A Deep Learning Approach. *Technologies* **2024**, *12*, 7. [CrossRef]
64. López-García, L.A.; Rodríguez-Cervantes, R.M.; Zamora-Martínez, M.G.; Esteban-Sosa, S.S. *Mis Manos que Hablan, Lengua de se nas para Sordos*; Editorial Trillas: Mexico City, México, 2008.
65. Cruz-Aldrete, M. *Gramática de la Lengua de Señas Mexicana*; El Colegio de México: Mexico City, Mexico, 2008.
66. Escobedo-Delgado, C.E., Ed. *Diccionario de Lengua de Señas Mexicana de la Ciudad de México*; INDEPEDI: Mexico City, Mexico, 2017.
67. Sánchez-Brizuela, G.; Cisnal, A.; de la Fuente-López, E.; Fraile, J.C.; Pérez-Turiel, J. Lightweight real-time hand segmentation leveraging MediaPipe landmark detection. *Virtual Real.* **2023**, *27*, 3125–3132. [CrossRef]
68. Hanke, T. HamNoSys—Representing Sign Language Data in Language Resources and Language Processing Contexts. In Proceedings of the LREC 2004, Workshop Proceedings: Representation and Processing of Sign Languages, Lisbon, Portugal, 26–28 May 2004; Streiter, O., Vettori, C., Eds.; European Language Resources Association (ELRA): Paris, France, 2004; pp. 1–6.
69. Rasheed, A.F.; Zarkoosh, M. Optimized YOLOv8 for multi-scale object detection. *J. Real-Time Image Process.* **2024**, *22*, 6. [CrossRef]
70. Wang, H.; Liu, C.; Cai, Y.; Chen, L.; Li, Y. YOLOv8-QSD: An Improved Small Object Detection Algorithm for Autonomous Vehicles Based on YOLOv8. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 1–16. [CrossRef]
71. Scikit-Learn. Accuracy Score. Available online: https://scikit-learn.org/stable/modules/model_evaluation.html#accuracy-score (accessed on 19 March 2025).
72. Scikit-Learn. Top-k Accuracy Score. Available online: https://scikit-learn.org/stable/modules/model_evaluation.html#top-k-accuracy-score (accessed on 19 March 2025).
73. Miah, A.S.M.; Hasan, M.A.M.; Shin, J. Dynamic Hand Gesture Recognition Using Multi-Branch Attention Based Graph and General Deep Learning Model. *IEEE Access* **2023**, *11*, 4703–4716. [CrossRef]
74. Khanam, R.; Hussain, M. YOLOv11: An Overview of the Key Architectural Enhancements. *arXiv* **2024**. arXiv:2410.17725.
75. Jocher, G.; Jing, Q.; Chaurasia, A. Ultralytics YOLO Docs: Image Classification. Available online: https://docs.ultralytics.com/tasks/classify/#models (accessed on 5 June 2025).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

sensors

MDPI

Article

A Multimodal Deep Learning Approach for Legal English Learning in Intelligent Educational Systems

Yanlin Chen [1,2], Chenjia Huang [1,3], Shumiao Gao [1], Yifan Lyu [1,4], Xinyuan Chen [1,5], Shen Liu [1,6], Dat Bao [2,*] and Chunli Lv [1,*]

1 China Agricultural University, Beijing 100083, China
2 Faculty of Education, Monash University, Melbourne, VIC 3800, Australia
3 Faculty of Humanities, China University of Political Science and Law, Beijing 102249, China
4 School of Informational Technology and Management, University of International Business and Economics, Beijing 100029, China
5 Law School, University of International Relations, Beijing 100091, China
6 National School of Development, Peking University, Beijing 100871, China
* Correspondence: dat.bao@monash.edu (D.B.); lvcl@cau.edu.cn (C.L.)

Abstract: With the development of artificial intelligence and intelligent sensor technologies, traditional legal English teaching approaches have faced numerous challenges in handling multimodal inputs and complex reasoning tasks. In response to these challenges, a cross-modal legal English question-answering system based on visual and acoustic sensor inputs was proposed, integrating image, text, and speech information and adopting a unified vision–language–speech encoding mechanism coupled with dynamic attention modeling to effectively enhance learners' understanding and expressive abilities in legal contexts. The system exhibited superior performance across multiple experimental evaluations. In the assessment of question-answering accuracy, the proposed method achieved the best results across BLEU, ROUGE, Precision, Recall, and Accuracy, with an Accuracy of 0.87, Precision of 0.88, and Recall of 0.85, clearly outperforming the traditional ASR+SVM classifier, image-retrieval-based QA model, and unimodal BERT QA system. In the analysis of multimodal matching performance, the proposed method achieved optimal results in Matching Accuracy, Recall@1, Recall@5, and MRR, with a Matching Accuracy of 0.85, surpassing mainstream cross-modal models such as VisualBERT, LXMERT, and CLIP. The user study further verified the system's practical effectiveness in real teaching environments, with learners' understanding improvement reaching 0.78, expression improvement reaching 0.75, and satisfaction score reaching 0.88, significantly outperforming traditional teaching methods and unimodal systems. The experimental results fully demonstrate that the proposed cross-modal legal English question-answering system not only exhibits significant advantages in multimodal feature alignment and deep reasoning modeling but also shows substantial potential in enhancing learners' comprehensive capabilities and learning experiences.

Keywords: visual and acoustic sensor integration; multimodal semantic fusion; human-centered intelligent education; vision–language–speech unified encoding

1. Introduction

With the rapid development of artificial intelligence and intelligent sensor technologies, language learning is gradually shifting towards a multimodal, interactive approach [1]. This shift is particularly important in legal English education. Legal English, being a highly specialized field, requires learners to master complex legal terminology, rigorous logical

structures, and extensive case analysis skills [2]. Moreover, legal English learning not only involves language proficiency but also requires learners to understand and express various forms of evidence and contexts [3]. In real-world legal work, case analysis often relies not only on textual materials but also on a vast array of evidence such as images, audio recordings, and other multimodal information [4]. This makes legal English learning content diverse and complex [5]. Traditional systems predominantly rely on text input and static exercises and lack systematic training in the semantic correlations between language and visual evidence, which makes it difficult for students to effectively apply legal language in real-life scenarios [6]. The traditional legal English teaching methods mainly rely on text-based learning and case analysis [7]. Through textbooks, legal texts, and case studies, students are able to acquire some basic legal knowledge and expressive skills [8]. However, this single-modality text-based approach has significant limitations. First, most of the learning content in traditional methods is based on written materials, with insufficient use of evidence types such as case images, videos, or audio [9]. In real legal practice, case judgments often depend on multiple sources of evidence, not just written text. Visual materials (e.g., photos of crime scenes, evidence images) and audio information (e.g., courtroom debates, witness testimonies) are essential components of legal analysis [10]. Traditional systems cannot effectively process these diverse input types, leaving learners unable to integrate and analyze these multimodal inputs when faced with complex cases [11]. Furthermore, traditional legal English education often remains static, with students only improving their language skills through reading and writing exercises, lacking opportunities for interaction in real-world contexts [12]. In actual legal work, learners need to understand complex legal texts and quickly respond in dynamic legal environments [13]. Lastly, traditional teaching methods offer limited interaction between students and the system, preventing learners from receiving timely feedback on their responses and lacking comprehensive evaluation of their responses. This results in difficulties in improving legal reasoning and language expression in practical applications [14].

The study of legal English represents a highly specialized task, requiring learners to simultaneously master semantic comprehension, evidentiary interpretation, and precise legal expression. These demands inherently involve multimodal information, including text, images, and speech, which poses significant challenges to traditional text-centered pedagogical methods due to their limited capacity for processing heterogeneous legal inputs such as courtroom imagery or spoken testimony. Although deep learning has opened new avenues for this domain [15], and recent advances in natural language processing (NLP), computer vision (CV), and speech recognition have enabled the integration of multimodal data into legal education [16], several limitations persist. Models such as BERT, CLIP, and wav2vec have demonstrated effectiveness in processing legal texts, aligning images with text, and transcribing speech into text, respectively [17,18]. However, their application in legal English instruction continues to encounter unresolved issues, including limited precision in cross-modal alignment, difficulty in conveying fine-grained semantic nuances across modalities, and the absence of real-time interactive feedback mechanisms. Prior research has highlighted the potential of cross-modal approaches, such as audio–text fusion for event detection [19], multimodal knowledge graph construction [20], and joint modeling of speech and text for emotion recognition [21,22], yet few studies have adequately addressed the distinctive challenges presented by legal discourse.

To address these challenges, a cross-modal question answering system for legal English, integrating image and audio sensor data, is proposed. This system leverages deep learning techniques across visual, textual, and speech modalities to construct an intelligent interactive platform applicable to diverse legal English education scenarios. It aims to facilitate the joint training of case comprehension and legal language expression in real-

istic legal contexts, as illustrated in Figure 1. The system acquires visual evidence and speech input through cameras and microphones, which are then processed via a structured three-stage pipeline. First, case images and legal questions are jointly encoded by a visual-language module. A CNN-Transformer hybrid network is employed to extract fine-grained visual features, which are fused with the Sentence-BERT-encoded question vectors through a cross-modal attention mechanism, yielding a unified semantic representation. Simultaneously, learners' spoken responses are processed by a speech recognition and semantic parsing module. This module utilizes a Transformer-based ASR model to transcribe the audio into text, which is subsequently interpreted by a diffusion Transformer into semantically enriched legal representations, ensuring coherence across temporal and modal contexts. Finally, the question–answer matching and feedback module integrates these multimodal features and employs a large-scale vision–language–speech fusion model to assess the quality of responses. A semantic alignment score is computed between the learner's input and the reference answer, and adaptive feedback is generated based on attention activations and similarity thresholds. Through this pipeline, raw sensor data are transformed into structured and interpretable legal language representations, enabling an interactive, context-aware legal English learning experience grounded in multimodal comprehension. The specific innovations introduced include:

1. Deep fusion of multimodal data: This system is the first to integrate data from image, speech, and text modalities for legal English question answering, extracting and understanding legal information from images and speech through deep learning models, breaking the limitations of traditional text-only input systems, and providing richer and more comprehensive learning materials.
2. Cross-modal question-answering system: The system not only understands the semantic relationship between images and text but also allows legal question answering through speech input, incorporating image evidence for comprehensive analysis. Users interact with the system through speech or image inputs, and the system provides real-time feedback and scoring based on the learner's responses, helping them better master the use of legal English.
3. Multimodal semantic fusion model: The system integrates CLIP for image–text semantic alignment, a Transformer-based ASR model for speech transcription and parsing, and Sentence-BERT for deep legal question encoding. On this basis, we propose a unified vision–language–speech embedding framework that supports fine-grained semantic integration across modalities. Furthermore, an affordance-driven masking mechanism and multi-level alignment strategy are introduced to enhance the model's focus on legally relevant information, enabling more accurate reasoning and feedback in complex legal scenarios.
4. Real-world scenario simulation and teaching feedback: By integrating image and speech sensors, the system simulates real courtroom scenarios or case analysis tasks, providing a more practical interactive learning experience. In interacting with the system, learners not only improve their legal English expression abilities but also enhance their legal thinking in real-world legal contexts.

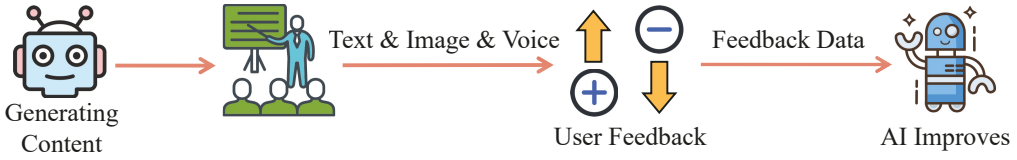


Figure 1. Illustration of the structure of multimodal feedback data composed of text, image, and voice signals that are collected and integrated to form the basis of sensor-enhanced intelligent education systems. The "+" symbol denotes the semantic integration of heterogeneous modalities, aiming to enrich learner understanding through complementary cues; the "−" symbol represents modal inconsistencies or interference, which the system resolves via alignment or correction strategies during multimodal fusion.

2. Related Work

2.1. Multimodal Semantic Understanding Technologies

Compared with traditional text-only understanding approaches, multimodal semantic understanding integrates image, text, and audio modalities, demonstrating significant advantages in information completeness, adaptability to heterogeneous inputs, and the realization of contextualized and interactive learning in educational scenarios. These advantages provide a more effective technical pathway for enhancing language transferability and semantic depth [23]. Recent advances in image–text joint learning (e.g., VisualBERT [24]) and speech–text alignment (e.g., wav2vec [25]) have substantially extended the applicability of natural language processing (NLP) in multimodal tasks [26]. These technologies enable systems to extract and align multimodal information, thereby improving the accuracy of semantic understanding [27]. For instance, VisualBERT extracts features and positional encodings from images and textual inputs using a pretrained CNN and a BERT encoder, respectively. The two modalities are then projected into a shared semantic space [28,29] and subsequently aligned via attention mechanisms, as formulated below:

$$\mathbf{v} = f_{\text{CNN}}(\text{Resize}(\mathbf{I})), \quad \mathbf{v} \in \mathbb{R}^{d_v}, \tag{1}$$

$$\mathbf{t} = f_{\text{BERT}}(\text{Tokenizer}(\mathbf{L})), \quad \mathbf{t} \in \mathbb{R}^{d_t}, \tag{2}$$

$$\mathbf{v}' = \mathbf{W}_v\mathbf{v} + \mathbf{b}_v, \quad \mathbf{t}' = \mathbf{W}_t\mathbf{t} + \mathbf{b}_t, \tag{3}$$

$$\mathbf{z} = \text{Attention}(\mathbf{v}', \mathbf{t}'), \quad \mathbf{z} \in \mathbb{R}^{d_z}, \tag{4}$$

where $\mathbf{I}$ and $\mathbf{L}$ denote the input image and text, respectively; f_{CNN} and f_{BERT} represent the image and text encoders; $\mathbf{W}_v$ and $\mathbf{W}_t$ are linear projection matrices; $\text{Attention}(\cdot)$ denotes the cross-modal attention alignment operation; and $\mathbf{z}$ is the final joint semantic representation. However, the static nature of features in VisualBERT limits its ability to model dynamic relations between visual details and textual semantics, rendering it less effective in handling image quality variability and semantic heterogeneity in real educational environments. CLIP further extends this framework by employing contrastive learning to optimize image–text matching within a shared feature space [30,31]. The contrastive loss is defined as:

$$\mathcal{L} = -\sum_i \log \frac{\exp(\mathbf{v}_i \cdot \mathbf{t}_i/\tau)}{\sum_j \exp(\mathbf{v}_i \cdot \mathbf{t}_j/\tau)}, \tag{5}$$

where τ is the temperature parameter and $\mathbf{v}_i$ and $\mathbf{t}_j$ represent the image and text feature vectors, respectively. This method enhances semantic consistency across modalities, thereby improving the accuracy of image–text retrieval tasks [32]. Despite its generalizability owing to large-scale training, CLIP lacks task-specific fine-tuning and thus exhibits limitations in recognizing domain-specific semantics such as legal terminology and evidentiary content. In the speech–text domain, wav2vec learns robust speech representations from unlabeled

audio data and maps raw waveforms into a feature space suitable for downstream transcription and semantic tasks [33,34], formalized as:

$$\mathbf{S} = \text{wav2vec}(\mathbf{X}), \tag{6}$$

where $\mathbf{X}$ denotes the input speech signal. For legal English learning scenarios, wav2vec enables accurate speech-to-text conversion and serves as a frontend for downstream semantic parsing pipelines [35]. However, it remains sensitive to speech variability such as accent and rate and provides only low-level acoustic features, which are insufficient for supporting high-level semantic reasoning and task-specific interpretation. Extended models such as SpeechBERT incorporate BERT's textual reasoning capabilities, with speech inputs to improve semantic alignment and parsing accuracy [36,37]. Nevertheless, these models are characterized by architectural complexity, high computational cost, and strong dependency on domain-specific corpora, which limits their scalability and responsiveness when deployed in practical educational systems.

In high-stakes educational domains such as legal English, models must not only achieve accurate semantic alignment but also ensure the traceability of reasoning paths and decision rationales. Current mainstream methods adopt end-to-end architectures with deeply nested neural layers, which, while powerful in representation, often exhibit "black-box" characteristics that obscure the influence of specific features on final outputs. Although some studies have introduced attention visualization and embedding space analysis to enhance transparency, these mechanisms remain insufficient to meet the demand for explicit explainability from instructors or learners. Therefore, the construction of multimodal models with structured interpretability and semantically layered representations is essential for ensuring the reliability and pedagogical applicability of such systems. This design motivation underpins the introduction of structural alignment mechanisms and semantic feedback modules in the proposed framework. Furthermore, in the "Results and Discussion" section, the proposed method is quantitatively compared with mainstream baselines across multiple metrics and error margins and statistical fluctuations are analyzed to validate its superiority in both accuracy and robustness.

2.2. Sensor-Supported Language Learning Systems

In the field of language learning, the integration of image and speech sensors has become a crucial tool for enhancing learning outcomes [38]. Speech scoring systems are among the most common applications, primarily using speech recognition technology to evaluate learners' pronunciation and provide feedback [39]. These systems typically assess pronunciation accuracy by comparing the standard pronunciation with the learner's pronunciation. The basic evaluation formula can be expressed as:

$$\text{Score} = \frac{1}{N}\sum_{i=1}^{N} \text{D}(y_i, \hat{y}_i), \tag{7}$$

where y_i represents the standard pronunciation, $\hat{y}_i$ represents the learner's pronunciation, N is the number of evaluated speech units, and $\text{D}(y_i, \hat{y}_i)$ represents the distance metric between the standard and learner's pronunciation. Common distance metrics include dynamic time warping (DTW) and Euclidean distance. This method allows the system to provide a score and specific improvement suggestions based on the differences between the learner's and the standard pronunciation [40]. Image reading practice platforms typically use image recognition technology to assist learners in recognizing and understanding the information in images [41]. For legal English learning, such platforms can be used to recognize images of crime scenes, evidence materials, etc., and provide learners with

related legal questions based on the image content [42]. This method helps learners better understand legal contexts and enhance their ability to perceive and judge case details. For example, the system automatically identifies and annotates key elements in the image, assisting learners in understanding the case background and evidence, leading to more accurate legal expression in subsequent question-answering tasks [43]. Although these sensor-supported systems have achieved good results in certain domains, they still have limitations in terms of task complexity and feedback granularity. For instance, in speech scoring systems, the evaluation mainly relies on pronunciation accuracy and fails to consider other language features such as grammar and semantics [44]. In image reading practice, although basic elements in images can be identified, the understanding of the underlying legal information in the image remains limited. Enhancing the adaptability of these systems in complex tasks, especially in the high-demand domain of legal English learning, remains a key challenge for future development [5].

2.3. Legal Language Education and Interactive Question-Answering Technologies

Legal language education, as a specialized form of language learning, requires learners to not only master basic legal terminology and grammar structures but also to possess strong legal reasoning and thinking skills [45]. Traditional legal English education methods typically focus on text reading and case analysis, but this approach often lacks a contextual learning experience, making it difficult to cultivate students' ability to handle cases in real legal scenarios [46]. As a result, interactive question-answering systems have been widely used in legal language education in recent years, especially in mock courtrooms and case analysis training, as these systems provide instant feedback to learners, helping them improve their legal understanding and expression abilities [47]. The core principle of interactive question-answering systems is based on NLP technology, which automatically understands learners' questions and generates appropriate answers [48]. Through multiple rounds of questioning, the system helps learners analyze cases, reason legal issues, and provide suitable legal advice. These systems typically establish a legal knowledge base to categorize and parse common legal issues [49]. With the development of deep learning technologies, many systems are now able to automatically understand complex legal problems and offer relevant solutions. For legal English learning, interactive question-answering systems help learners practice language skills in real legal contexts and improve their ability to handle complex legal issues [50]. However, existing interactive question-answering systems still face some challenges in handling complex legal issues. In particular, during legal English learning, systems need to understand not only legal terminology and case background but also multimodal information such as evidence materials [51]. Traditional question-answering systems often focus on a single modality (e.g., text) and lack effective handling of other sensory inputs such as images and speech [2]. Therefore, introducing multimodal information into legal question-answering systems and achieving effective alignment between modalities is a critical direction for improving system performance. The multimodal legal English question-answering system proposed in this paper is designed to address these issues by integrating image, speech, and text information, aiming to provide more comprehensive and accurate support for legal English learning. By incorporating deep learning models, the system can process the semantic relationships between different modality data, offering learners a richer learning experience.

3. Materials and Methods

3.1. Data Collection

The construction and use of datasets in AI-assisted legal applications—particularly those involving judicial or institutional scenarios—must adhere to emerging legal and

regulatory frameworks that govern the access, provision, and processing of source data. The dataset presented in this work was developed in accordance with responsible data handling principles, including transparency, authorized public access, and the exclusion of sensitive or private content. These considerations align with ongoing discussions in regulatory frameworks such as the U.S. Federal Rules of Evidence and the EU's General Data Protection Regulation (GDPR), which emphasize lawful sourcing, traceability, and privacy protection in AI-based legal systems.

A dataset targeting the multimodal legal English question-answering task was constructed, encompassing data from three modalities: images, text, and audio. The image data were primarily collected from publicly available online platforms, including the United States court record website (CourtListener), the well-known legal resource platform FindLaw, and Wikimedia Commons. Specifically, the collected image types include case scene photographs, evidence item images, and scanned copies of documents and contracts, with each category containing between 1000 and 2000 samples, as shown in Table 1. To ensure the quality and semantic relevance of the images, a uniform resolution not lower than 1024 × 768 pixels was adopted during the collection process, prioritizing images with clear annotations, complete scenes, and significant legal evidentiary value. All images were manually filtered and preliminarily annotated after collection, with annotations covering case types (such as criminal, civil, and intellectual property), image themes (such as evidence items and accident scenes), and key information regions, thereby ensuring high usability and contextual relevance within legal scenarios. In terms of data processing, all samples underwent a standardized preprocessing pipeline involving resolution normalization, noise filtering, and semantic-preserving augmentation strategies tailored to legal visual content. The quality of the samples was evaluated through both technical and semantic criteria, including clarity, annotation completeness, and relevance to legal contexts. To address the limitation in data volume and ensure robustness in diverse learning scenarios, an open-source dataset, MultimodalQA, was further incorporated as an auxiliary resource for experimentation. This dataset complements the in-house collection by providing a broader distribution of multimodal question–answer pairs, thereby enabling more comprehensive training and generalization analysis. Regarding textual data, the questions and standard answers were compiled from public case commentaries and question–answer style analyses published by the United States Supreme Court, FindLaw case database, and Harvard Law Review. Approximately 1500 pairs of questions and answers were collected, covering major fields such as criminal law, contract law, tort law, and intellectual property law. It should be noted that these textual question–answer pairs were constructed independently of visual or auditory modalities and thus do not include any associated images or video materials. During text processing, the characteristics of the legal question–answer genre were strictly followed, ensuring that the questions exhibited appropriate complexity (e.g., involving multi-step reasoning, fact summarization, and legal application), while the answers remained organized, logically coherent, and reflective of standard legal expression norms. Additionally, to enhance the educational applicability of the corpus, certain texts were moderately simplified, maintaining professionalism while lowering language complexity to facilitate learner comprehension and response. Audio data were collected through simulated learning scenarios, mainly using Blue Yeti noise-reduction microphones and associated acoustic sensor devices. The recording subjects included both native English speakers and advanced English learners, with a balanced gender distribution and an age range spanning from 20 to 55 years. To ensure diversity, participants exhibited varied accents, including American, British, and Southeast Asian English. Each audio sample corresponded to one question, with a duration controlled between 10 and 90 s. Recordings were conducted in a quiet, echo-free environment, and a

consistent sampling rate of 16 kHz at 16-bit resolution was applied. To further enhance the dataset's naturalness and robustness, the speech samples encompassed variations in speaking speed (ranging from slow to fast-paced delivery), intonation patterns (from flat to expressive), and degrees of formality in legal expression. Additionally, mild background noise simulations (e.g., soft environmental sounds) were introduced to simulate realistic acoustic conditions. In total, approximately 1200 high-quality audio recordings were compiled, providing a solid foundation for training and evaluating multimodal models under diverse and practical speech scenarios.

The current dataset is closely aligned with the terminological conventions, argumentative structures, and case organization typical of the Anglo-American common law system, which facilitates effective modeling within this legal context. However, such alignment may limit the system's applicability to other legal traditions, such as civil law systems, where legal reasoning and textual formats differ significantly. Moreover, since the speech data were primarily collected from native English speakers, the ASR component may exhibit elevated error rates when processing inputs from non-native learners, thereby reducing its robustness and inclusiveness across diverse learner populations.

Table 1. Data quantity for different types of multimodal legal English resources.

Data Type	Quantity
Case Scene Images (CourtListener)	1247
Evidence Document Images (FindLaw)	1372
Legal Document Scans (Wikimedia Commons)	1098
Legal QA Text Pairs (Supreme Court/FindLaw/Harvard Law Review)	1520
Learner Response Audios (Recorded by Blue Yeti)	1214

3.2. Data Preprocessing and Augmentation

In multimodal learning systems, data preprocessing and augmentation are crucial steps for improving model performance and robustness, particularly in image and audio data processing. Appropriate preprocessing and augmentation techniques can effectively enhance the model's adaptability to input data. For image and audio data, preprocessing primarily aims to clean and normalize raw data, while data augmentation generates variations of the data to improve the model's ability to generalize to different input conditions. In this study, all preprocessing procedures followed widely adopted standards in the computer vision and speech processing communities. Image data were handled using OpenCV-based pipelines, incorporating resolution normalization ($\geq$1024 $\times$ 768), denoising via median filtering, and geometric transformations constrained by semantic region masks. Audio preprocessing adhered to standard speech recognition settings, including a 16 kHz sampling rate, 16-bit depth, and silence removal using amplitude-based thresholding. Feature extraction and augmentation were implemented using industry-standard libraries such as librosa and PyDub. These choices ensure that the data fed into the model meet both technical consistency and reproducibility requirements.

3.2.1. Image Data Preprocessing and Augmentation

Image data preprocessing and augmentation are essential in CV tasks because the quality and content of images directly affect the model's learning performance. Common image preprocessing operations include edge detection, resolution standardization, and noise filtering. These steps help the model better extract meaningful information from the image while reducing the interference from irrelevant factors. Edge detection is a commonly used technique in image preprocessing aimed at highlighting important features in an image, such as object contours or edges. Common edge detection algorithms

include Sobel operator and Canny edge detection. The Sobel operator is a gradient-based method used to detect edges in images. The computation process can be expressed by the following formulas:

$$G_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \otimes I, \tag{8}$$

$$G_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} \otimes I, \tag{9}$$

where G_x and G_y represent the horizontal and vertical gradients of the image, $\otimes$ denotes the convolution operation, I represents the input image, and the kernels are manually defined based on the classical Sobel operator, as this step is used for fixed edge enhancement during preprocessing rather than trainable feature extraction. By applying Sobel convolution to the image, edge information is obtained, enhancing the prominence of important structures in the image. Resolution standardization is another common preprocessing step. In multimodal learning systems, image inputs may come from different sources, leading to variations in resolution and size. To ensure consistent processing of image data, it is necessary to resize all images to a standard resolution. For example, for an input image $\mathbf{I}$, it can be resized to a fixed size $\mathbf{I}'$ using interpolation methods:

$$\mathbf{I}' = \text{Resize}(\mathbf{I}, (w, h)), \tag{10}$$

where $\text{Resize}(\mathbf{I}, (w, h))$ refers to resizing the input image $\mathbf{I}$ to dimensions (w, h). Common interpolation methods include nearest-neighbor interpolation and bilinear interpolation. Noise filtering refers to removing irrelevant noise from images, typically accomplished using median filtering or mean filtering. Noise can affect the extraction of image features, leading to decreased model performance. Median filtering is a commonly used denoising method that replaces the target pixel with the median of its neighboring pixels to reduce the impact of noise. Its computation can be expressed as:

$$I_{\text{new}}(x, y) = \text{median}(\{I(x + i, y + j)\}_{i=-k, j=-k}^{k}), \tag{11}$$

where $I_{\text{new}}(x, y)$ represents the denoised pixel value, $\{I(x + i, y + j)\}$ denotes the neighboring pixels around the target pixel, and k is the size of the filter. Through this method, noise in the image can be effectively suppressed. In addition to these basic image preprocessing methods, data augmentation techniques such as rotation, cropping, flipping, and scaling are employed to generate diverse image variants, enhancing the model's adaptability to different scenarios. To ensure semantic integrity during augmentation, a region protection mechanism is applied based on annotated evidential areas, where transformations are restricted to non-critical regions to avoid occluding or distorting legally relevant content. Furthermore, cropping operations are guided by a saliency-aware and boundary-constrained strategy, which preserves the interpretability of key visual elements. A consistency validation pipeline is also integrated before and after augmentation to assess both image quality and content fidelity, ensuring that the generated data remain legally interpretable and applicable for downstream model training. For example, image rotation helps the model better learn rotation-invariant features. For an image $\mathbf{I}$, rotation can be represented as:

$$\mathbf{I}' = \text{Rotate}(\mathbf{I}, \theta), \tag{12}$$

where θ is the rotation angle and $\text{Rotate}(\mathbf{I}, \theta)$ denotes the new image $\mathbf{I}'$ obtained by rotating the image $\mathbf{I}$ by angle θ. Additionally, cropping helps the model focus on important

information in the image by randomly cropping parts of the image to increase the diversity of training samples:

$$\mathbf{I}' = \text{Crop}(\mathbf{I}, (x_1, y_1), (x_2, y_2)), \tag{13}$$

where (x_1, y_1) and (x_2, y_2) represent the top-left and bottom-right coordinates of the cropping area, respectively.

3.2.2. Audio Data Preprocessing and Augmentation

Audio data preprocessing and augmentation are also key steps in improving the model's robustness and performance. Audio signals typically consist of a series of time-sequence data, and it may be difficult to extract meaningful features by directly processing these raw data. Therefore, several preprocessing operations are necessary. Common audio preprocessing methods include silence removal, speech rate normalization, and pitch enhancement. Silence removal is an important step in audio preprocessing, aiming to remove silent sections from the audio signal. Silent sections in audio can occupy storage space and potentially impact the model's training efficiency. Therefore, by detecting silent portions and removing them, the quality of audio data can be effectively improved. Silence removal is typically performed by setting a threshold α to determine if the audio signal is silent:

$$\text{Silence}(\mathbf{X}, \alpha) = \{t \mid |\mathbf{X}(t)| < \alpha\}, \tag{14}$$

where $\mathbf{X}(t)$ represents the amplitude of the audio signal at time t, α is the silence threshold, and $\text{Silence}(\mathbf{X}, \alpha)$ denotes the audio segments that are silent. By removing silent sections, ineffective information in the audio can be eliminated. Speech rate normalization refers to adjusting the speech rate of audio to a certain range so that speech inputs at different rates can be uniformly processed. Speech rate normalization is typically achieved by stretching or compressing the duration of the audio signal. Assuming the duration of the audio signal is T and the target duration after speech rate normalization is T', audio resampling can be performed by the following formula:

$$\mathbf{X}'(t) = \mathbf{X}\left(\frac{t}{T} \cdot T'\right), \tag{15}$$

where $\mathbf{X}'(t)$ represents the normalized audio signal, T is the original duration, and T' is the target normalized duration. Speech rate normalization eliminates the impact of varying speech rates, ensuring consistency in the audio data. Pitch enhancement is a method used to change the pitch of the speech by adjusting its frequency, enabling the model to better adapt to varying speech characteristics. Pitch enhancement is typically achieved by modifying the frequency components of the audio signal, either stretching or compressing the frequency. Let $F(\mathbf{X})$ represent the frequency components of the audio, and pitch enhancement can be represented as:

$$F'(\mathbf{X}) = \alpha F(\mathbf{X}), \tag{16}$$

where α is the pitch adjustment factor, $F(\mathbf{X})$ represents the frequency components of the original audio signal, and $F'(\mathbf{X})$ represents the frequency components after pitch enhancement. By enhancing the pitch, the pitch of the audio signal is adjusted, helping the model adapt to more variations in speech inputs. Spectral simulation extension is an important method in audio data augmentation. By performing spectral simulation extension on the original audio signal, diverse audio samples can be generated to enhance the model's robustness. Spectral simulation usually involves randomly transforming the audio's spectrogram, such as perturbing or scaling the magnitude of the spectrogram,

thereby generating new audio samples. This approach allows the model to adapt to more complex and diverse audio inputs, improving its performance in real-world applications.

3.3. Proposed Method

The proposed cross-modal legal English question-answering system is designed with the primary methodological objective of enabling accurate semantic alignment and interactive reasoning across visual, textual, and auditory modalities in legal educational contexts. Specifically, the system aims to (1) bridge the gap between perceptual evidence and legal language expression, (2) enhance the learner's ability to comprehend and respond to complex multimodal case scenarios, and (3) provide interpretable, adaptive feedback to support progressive language acquisition. To achieve these goals, the architecture is composed of three sequential modules: a visual-language encoding module, a speech recognition and semantic parsing module, and a question–answer matching and feedback module. The processed case images and textual questions are first input into the visual-language encoding module, where deep feature extraction and normalization are performed to generate a unified visual-language semantic representation. Learners' speech, collected via acoustic sensors, is then input into the speech recognition and semantic parsing module to be transcribed into text and to extract key information, which is subsequently aligned bidirectionally with the case semantics. Finally, in the question–answer matching and feedback module, similarity modeling and matching scoring are conducted based on multimodal fusion vectors, generating detailed feedback to assist language learning optimization, as shown in Figure 2. Additionally, in order to provide a clearer understanding of the system's architecture and each component's role, Table 2 summarizes the key modules, associated input modalities, underlying models, primary functions, and output formats.

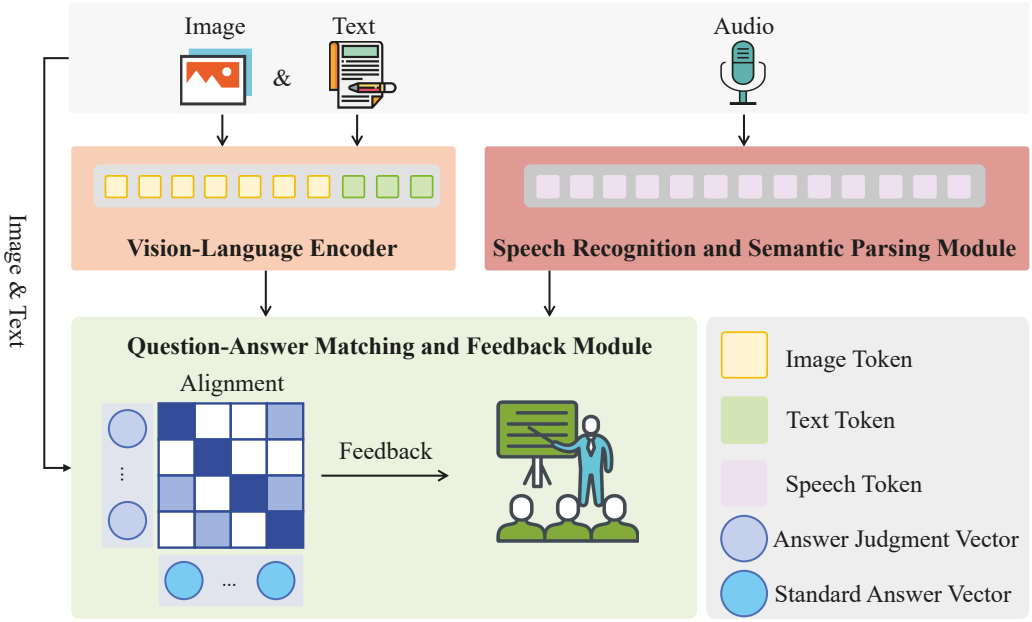


Figure 2. Overall architecture of the proposed cross-modal legal English question-answering system. The system sequentially integrates a visual-language encoding module, a speech recognition and semantic parsing module, and a question–answer matching and feedback module.

Table 2. Overview of module functions in the proposed cross-modal legal QA system.

Module Name	Input Modalities	Core Models Used	Primary Function	Output Format
Visual-Language Encoding Module	Case image, question text	ResNet-50, SBERT, Multi-head Attention	Extracts deep visual and textual features and generates a unified semantic representation	1024-dim fused vector
Speech Recognition and Semantic Parsing	Learner's spoken response	Transformer ASR, Diffusion Transformer	Converts audio to text and extracts legal semantic structure aligned with visual-question context	Sequence of 512-dim vectors
Question–Answer Matching and Feedback	Fused visual, text, and speech features	ViT-B/16, OPT-1.3B LLM, Feedback Generator	Performs multimodal answer matching and generates adaptive, interpretable feedback for learners	Matching score and feedback

3.3.1. Visual-Language Encoding Module

The visual-language encoding module is designed to achieve joint semantic understanding of case images and legal question texts, thereby generating a unified representation that serves as the foundation for subsequent cross-modal question answering. As illustrated in Figure 3, the input image is first processed through a convolutional backbone network to extract hierarchical spatial features. These features are subsequently segmented into patch tokens and embedded via a Transformer Encoder to capture global contextual relationships. In parallel, the legal question text is encoded using the Sentence-BERT model to obtain a sentence-level semantic vector that preserves the logical structure and linguistic nuances of legal discourse. To bridge the modality gap, visual and textual embeddings are projected into a shared semantic space, where cross-modal alignment is performed using a weighted attention mechanism. This process enables the integration of evidentiary visual cues with legal language semantics, resulting in a compact and interpretable multimodal feature representation. Specifically, the ResNet-50 architecture is employed as the backbone network in the image branch. ResNet-50 was chosen due to its strong balance between depth and computational efficiency, as well as its proven performance in capturing hierarchical features from structured and high-detail legal evidence images. Input case images are resized to 224×224 resolution and passed through four convolutional blocks, each containing a convolutional layer (with kernel size 3×3), batch normalization (BatchNorm), and ReLU activation units, progressively expanding the feature dimensions from 64 to 2048. After generating the convolutional feature maps, a multi-head self-attention mechanism (Multi-Head Attention) is introduced to meet the high-level semantic modeling requirements of cross-modal tasks. Each image is partitioned into 14×14 patch tokens and encoded using a Transformer Encoder, where each token has a dimensionality of $D = 512$, with a depth of 6 layers and 8 attention heads per layer.

For the text branch, Sentence-BERT (SBERT) is utilized as the encoder, producing a $D = 768$ dimensional semantic vector for each legal question text via a pretrained Transformer network, as it is specifically designed to generate high-quality sentence-level embeddings while maintaining contextual semantic coherence—an essential property for capturing the intricate logical structures and multi-clause dependencies commonly found in legal discourse. To mitigate the dimensionality discrepancy between modalities and facilitate subsequent fusion, the text vector is projected to 512 dimensions via a linear transformation, formulated as $\mathbf{t}' = \mathbf{W_t}\mathbf{t} + \mathbf{b_t}$, where $\mathbf{W_t} \in \mathbb{R}^{512\times768}$ is a learnable weight matrix and $\mathbf{b_t}$ is the bias term. Visual and textual features are matched and aligned through a feature fusion layer. A weighted alignment strategy is adopted, whereby similarity

scores $s_i = \cos(v_i, t')$ are computed between each image token sequence $v_1, v_2, \ldots, v_n$ and the projected text vector t', followed by Softmax normalization to obtain a weight distribution $p_i = \frac{\exp(s_i)}{\sum_j \exp(s_j)}$. The fused visual semantic representation is calculated as a weighted sum $\mathbf{v}fused = \sum i = 1^n p_i v_i$, where $n = 196$ denotes the number of image tokens. The final visual-language joint representation is obtained by concatenating the fused visual semantics and the text semantics, expressed as $z = [\mathbf{v}_{fused}; t']$, resulting in a 1024-dimensional unified encoded vector. To further enhance the module's focus on key information regions, an affordance masking mechanism is introduced at the image feature extraction stage. The affordance masking mechanism prioritizes evidentiary image regions directly relevant to the current legal question, improving semantic integrity during fusion. Specifically, based on the content of the current question text, a mask matrix $\mathbf{M} \in 0, 1^{14\times14}$ is predicted to identify potential attention regions, retaining only the features where the mask value is 1 for subsequent processing. The mask prediction is achieved through a lightweight convolutional network, conditioned on the question text encoding vector. This mechanism effectively suppresses irrelevant background noise and non-target objects within the case images, thereby enabling the model to concentrate more accurately on evidential and reasoning-related visual clues. During training, the optimization objective of the visual-language encoding module is to maximize the similarity between positive image–text pairs (I, Q) while minimizing the similarity for negative pairs (I, Q') or (I', Q). The loss function is based on an InfoNCE contrastive learning loss, defined as:

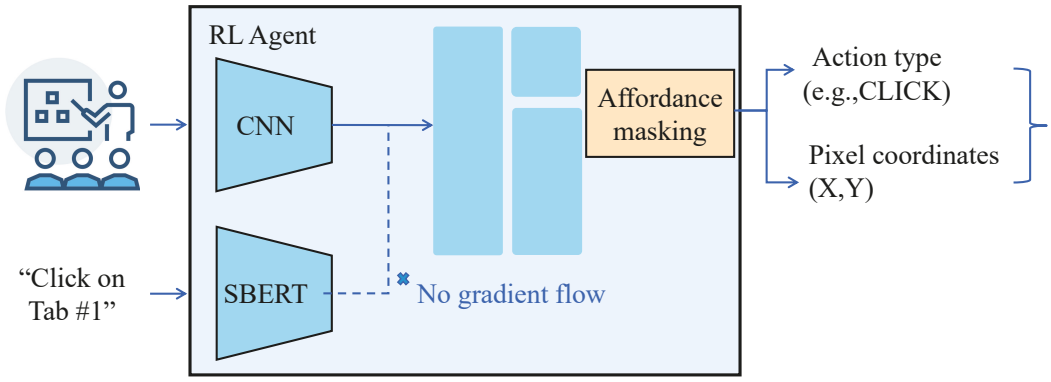


Figure 3. Illustration of the visual-language encoding module. The module extracts regional image features using a CNN and encodes textual semantics with a Sentence-BERT (SBERT) model to achieve deep multimodal correlation.

$$\mathcal{L}contrastive = -\log \frac{\exp(\cos(z_I, z_Q)/\tau)}{\sum z' \in \mathcal{Z} \exp(\cos(z_I, z')/\tau)}, \tag{17}$$

where z_I and z_Q represent the encoded vectors for the image and text inputs respectively, τ denotes the temperature parameter, and $\mathcal{Z}$ refers to the set of all text encodings within the current batch. Through contrastive learning optimization, the model effectively brings correct image–text pairs closer in the shared semantic space while pushing apart mismatched pairs, thereby enhancing the cross-modal alignment capability of the encoding module. The design of the visual-language encoding module provides significant advantages for the targeted task. The combination of convolutional features and Transformer encoding enables the model to capture both local detail sensitivity and global contextual reasoning ability, allowing accurate association of critical case information with the corresponding legal questions. The affordance masking mechanism further strengthens the module's focus on relevant regions, substantially improving the generalization performance across diverse legal case images. Moreover, the unified dimensionality design of the cross-modal encoding simplifies subsequent semantic matching operations and reduces inference latency,

ensuring the responsiveness and interactivity of the system. Mathematically, the weighted feature fusion and contrastive learning training strategy allow the visual-language encoding module to form a more compact matching distribution in the feature space, thereby improving the matching accuracy and robustness under complex legal scenarios.

3.3.2. Speech Recognition and Semantic Parsing Module

The speech recognition and semantic parsing module was designed to accurately extract textual information from learners' speech input and further capture legal-related semantic features, thereby providing high-quality semantic representations for subsequent question–answer matching and feedback modules. As shown in Figure 4, the raw audio signals collected by acoustic sensors are first subjected to preprocessing steps including noise filtering, silence removal, and time–frequency feature extraction. These extracted features are then fed into a transformer-based automatic speech recognition (ASR) model to generate preliminary textual transcriptions. The ASR model was constructed based on a Transformer architecture composed of $N = 12$ stacked Transformer Encoder layers, each consisting of a multi-head self-attention mechanism, a feedforward network, and a layer normalization module. The input features were initially downsampled via convolutional layers, where the raw audio feature dimensions were $(T = 160, F = 80, C = 1)$, with T representing time steps, F denoting frequency channels, and C corresponding to single-channel grayscale signals. After three convolutional blocks, each with a 3×3 kernel size and a 2×2 stride, the channel dimensions were progressively expanded to 32, 64, and 128, with the final output feature size reduced to $(T/8, F/8, 128)$.

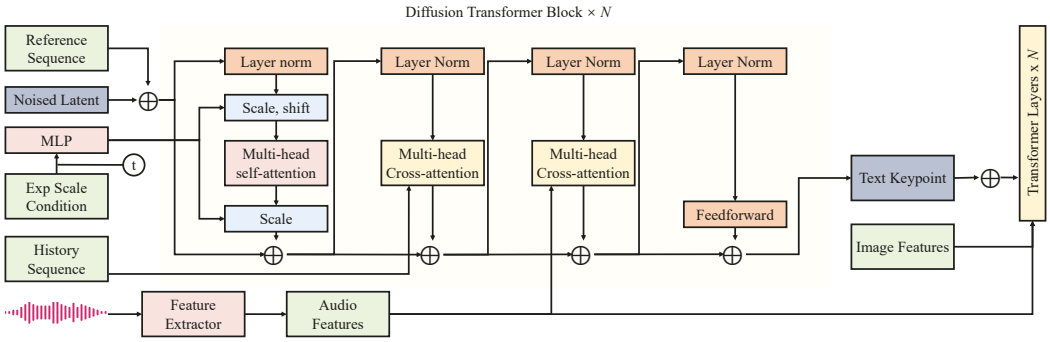


Figure 4. Illustration of the speech recognition and semantic parsing module. The module initially transcribes speech signals into audio features using an ASR model, incorporating inputs from case images, body keypoints, and hand coefficients to extract multisource features.

The convolutional features were then flattened and augmented with positional encoding before being input into the Transformer Encoder. Each self-attention layer utilized $H = 8$ attention heads, with each head having key, query, and value vector dimensions of $d_k = d_q = d_v = 64$ and a feedforward layer with $d_{ff} = 2048$ hidden units and a GELU activation function. The Transformer Encoder outputted a textual feature sequence $\mathbf{S} = s_1, s_2, \ldots, s_n$, where $s_i \in \mathbb{R}^{512}$ represents the textual feature vector at the i-th time step. Due to the inherent limitations of conventional ASR architectures in capturing long-range semantic dependencies and maintaining contextual coherence in complex legal discourse, a series of extended cross-modal diffusion Transformer modules were introduced following the ASR model. These modules employ a multi-step semantic reconstruction mechanism, conditioned on historical semantic trajectories, to dynamically integrate contextual information while preserving the original speech features. This design enhances the model's capacity to identify implicit logical structures and key legal terminology within extended spoken statements. The diffusion Transformer comprised $N = 8$ diffusion blocks, each containing a multi-head cross-attention module, a multi-head self-attention module, a feed-

forward network, and a layer normalization module. The cross-modal attention utilized the current frame's textual features as queries, with historical reference sequences serving as keys and values, thereby guiding the dynamic incorporation of historical semantic contexts into the current understanding process. After each cross-modal interaction, residual connections preserved the original semantic trajectory, while adaptive scale and shift operations adjusted feature magnitudes to ensure semantic stability during diffusion. To address feature consistency across different scales, a consistency fusion mechanism was introduced. After each block's output, feature normalization and rescaling were performed, followed by minimizing the discrepancy with the initial reference features. The consistency fusion loss $\mathcal{L}_{consistency}$ was defined as:

$$\mathcal{L}consistency = \frac{1}{n}\sum i = 1^n |s_i - r_i|_2^2, \tag{18}$$

where s_i denotes the i-th output textual feature, r_i represents the corresponding historical reference feature, and n is the total number of time steps. The consistency fusion loss ensures that semantic drift is minimized during multi-step reasoning, thereby preserving the core legal meanings in speech responses. Moreover, during multimodal joint reasoning involving visual evidence and textual answers, the semantic parsing module was aligned with the outputs of the visual-language encoding module. A cross-modal alignment loss $\mathcal{L}_{align}$ was introduced, defined as:

$$\mathcal{L}align = -\frac{1}{n}\sum i = 1^n \log \frac{\exp(\cos(s_i, v_i)/\tau)}{\sum_{j=1}^{n} \exp(\cos(s_i, v_j)/\tau)}, \tag{19}$$

where s_i denotes the parsed textual feature at time step i, v_i represents the corresponding encoded vector of the case image–text pair, τ is the temperature parameter, and $\cos(\cdot, \cdot)$ indicates cosine similarity. By maximizing the similarity between correct text–image pairs and suppressing incorrect ones, the module enhanced feature discriminability during multimodal information fusion, thereby improving the precision of subsequent question–answer matching. The total loss function $\mathcal{L}_{total}$ for this module was formulated as a weighted sum of the individual losses:

$$\mathcal{L}total = \lambda_1 \mathcal{L}ASR + \lambda_2 \mathcal{L}consistency + \lambda_3 \mathcal{L}align, \tag{20}$$

where $\mathcal{L}_{ASR}$ is the standard cross-entropy loss for speech recognition and $\lambda_1, \lambda_2, \lambda_3$ are weighting coefficients set to $0.6, 0.2, 0.2$, respectively, to achieve an optimal balance between recognition accuracy and cross-modal consistency. Through the above module design, the system ensured high accuracy in speech recognition while further enhancing the quality of the extracted legal semantic structures. By deeply collaborating with the visual-language encoding module, efficient alignment between case facts and language expression in the cross-modal legal English question-answering scenario was realized. Particularly, the integration of the diffusion Transformer and consistency fusion strategies significantly improved the system's robustness and adaptability to complex inputs, such as accent variations, speech rate fluctuations, and case diversity, thereby ensuring broad applicability and high performance in practical educational settings.

3.3.3. Question–Answer Matching and Feedback Module

The question–answer matching and feedback module was designed to achieve deep semantic alignment between learners' spoken responses, case images, and legal question texts, while generating personalized and targeted feedback to enhance the interactivity and instructional effectiveness of the cross-modal legal English question-answering system.

As shown in Figure 5, the module comprises a vision encoder, a projection layer, a large language model (LLM), and a normalization and inference layer. Various data sources, including images, multimodal fused features (outputs from the visual-language encoding and speech parsing modules), textual questions, and task instructions, were preprocessed and embedded through specific paths before being fused, matched, and output.

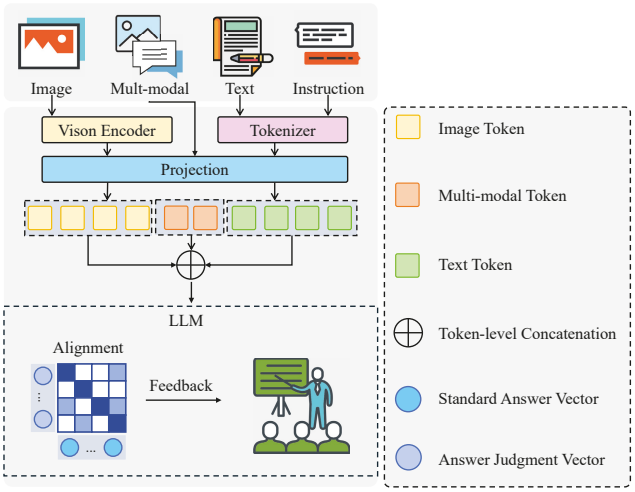


Figure 5. Illustration of the question–answer matching and feedback module. The module aims to compute the semantic alignment between learner responses and reference answers by integrating multimodal inputs, including case images, legal questions, spoken responses, and instructional prompts. Case images are processed via a Vision Transformer Encoder, while textual and speech features are embedded into a shared space through a projection mechanism. The resulting fused vector is fed into a large language model for contextual inference. A normalized similarity score is then calculated against the reference vector, guiding the generation of interpretable, template-based feedback informed by attention activations.

The input components included case images, multimodal fusion features, textual question content, and prompt instructions. Case images were processed through a Vision Encoder based on the ViT-B/16 (Vision Transformer Base) architecture, which was selected for its superior capability in modeling long-range dependencies and global context—essential for capturing complex evidentiary relationships in legal imagery. This architecture consists of 12 transformer blocks with a hidden width of 768 and 12 self-attention heads, each with a dimension of 64, maintaining an overall feature dimension of 768. Images were partitioned into 16×16 patches, each linearly projected into a 768-dimensional vector and then passed through the Transformer to model global relationships. After encoding, a visual semantic vector $\mathbf{v} \in \mathbb{R}^{768}$ was produced. Correspondingly, textual content and instructions were tokenized and embedded into 768-dimensional vectors, ensuring feature space consistency with the visual vectors. To effectively integrate image, text, and other modality inputs within a unified feature space, a projection layer was introduced after the Vision Encoder, defined by the linear mapping:

$$\mathbf{v}' = \mathbf{W}_p \mathbf{v} + \mathbf{b}_p, \tag{21}$$

where $\mathbf{W}_p \in \mathbb{R}^{768 \times 768}$ and $\mathbf{b}_p$ is a learnable bias term. This projection operation retained the feature dimension while adjusting the distribution of visual features to better align with the input requirements of the large language model. After projection, all modality inputs were concatenated within the embedding space to form a complete multimodal input sequence, serving as the basis for subsequent reasoning. The multimodal input sequence was fed into a pretrained large language model based on the OPT-1.3B architecture, which was selected as the backbone after comparative evaluation of multiple open-source models of varying parameter scales, as it provides a favorable balance between inference capability and com-

putational efficiency—consisting of 24 Transformer blocks, each with a width of 2048 and 32 self-attention heads, each head with a dimension of 64. Through layer normalization and multi-head attention mechanisms, the model implicitly learned cross-modal relationships during inference. The large language model, via context modeling, established high-level semantic correspondences between case images, question texts, and learners' responses and outputted a final answer judgment vector $\mathbf{o} \in \mathbb{R}^{2048}$. To evaluate the matching degree between learners' answers and standard answers, a normalization and matching layer was introduced. The matching process employed an inner product similarity metric, where the output vectors were normalized and the cosine similarity with the standard answer vector $\mathbf{o}_{std}$ was computed as follows:

$$\mathrm{sim}(\mathbf{o}, \mathbf{o}std) = \frac{\mathbf{o} \cdot \mathbf{o}std}{|\mathbf{o}||\mathbf{o}_{std}|}. \tag{22}$$

The matching score was then input into the feedback generation submodule. Based on the score and local attention activations, personalized feedback texts were generated following a template-based strategy. To further enhance interpretability, the feedback module integrates an attention visualization mechanism and modular template design to support the traceability of decision paths. Learners can review which visual or linguistic components were emphasized during answer generation, increasing the transparency of the assessment and enabling more targeted improvement. Although the system is built upon predefined structures, the feedback content is not strictly fixed. Instead, it dynamically selects and assembles appropriate prompt phrases from a template pool based on the question–answer matching score and attention-weighted focus regions, embedding specific fragments from the learner's response to produce targeted and context-aware feedback. Specifically, when the similarity exceeded the threshold $\theta = 0.85$, positive feedback was provided, encouraging learners to attempt more complex reasoning. When the similarity fell within the range $0.6 < \mathrm{sim} < 0.85$, the system activated partially correct templates and, by identifying missing or misaligned elements through attention distribution, offered interpretable suggestions for refinement. In cases where the similarity dropped below 0.6 and the learner's response deviated significantly from expectations, the system prompted a re-evaluation of key case elements by presenting scaffolded instructional cues, including decomposed sub-questions and reflective prompts, thereby enhancing the openness and guidance of the feedback mechanism. To improve feedback consistency and contextual coherence across multiple rounds of interaction, a cumulative feedback loss $\mathcal{L}_{feedback}$ was designed, defined as:

$$\mathcal{L}feedback = \frac{1}{N}\sum i = 1^{N}(1 - \mathrm{sim}(\mathbf{o}i, \mathbf{o}std, i))^2, \tag{23}$$

where N represents the number of historical interaction rounds, $\mathbf{o}i$ is the output vector at the i-th round, and $\mathbf{o}std, i$ is the corresponding standard answer vector. This loss encouraged the model to maintain output consistency across interactions, avoiding feedback fluctuations caused by local feature noise, thereby providing learners with a stable and coherent learning experience. The design of this module offers substantial advantages for the targeted task. On one hand, by introducing the Vision Encoder and projection mechanism, deep integration of case image information and textual semantics within the feature space was achieved, requiring learners to not only comprehend text but also accurately reason based on visual evidence. On the other hand, by leveraging the context modeling capabilities of the large language model, complex implicit cross-modal correspondences within case reasoning processes were effectively captured, enhancing the model's understanding of complex responses. Furthermore, the introduction of the normalization-based

matching strategy and cumulative feedback loss ensured the precision of answer matching and the stability of feedback generation, significantly enhancing learners' comprehensive application abilities and learning outcomes in authentic legal English scenarios.

4. Results and Discussion

4.1. Hardware and Software

The hardware platform of this system includes a high-performance workstation equipped with the latest NVIDIA A100 GPU (Nvidia, Santa Clara, CA, USA), which provides powerful computing capabilities and can efficiently support deep learning model training and inference. The workstation also features 32GB of memory and a 1TB solid-state drive, providing ample storage and processing space to handle large-scale dataset loading and real-time processing requirements. For input devices, we use a high-definition camera and a high-sensitivity microphone, with the former capturing high-quality legal case images and the latter accurately capturing a learner's speech input. Additionally, the system is equipped with multiple external hard drives for data backup and storage, ensuring data security and the efficient operation of the system. In terms of software, the system is developed in a Linux-based environment, specifically using Ubuntu 20.04 LTS to ensure stability and compatibility. The deep learning framework used is primarily PyTorch 2.1, with TensorFlow 2.14 utilized for the optimization and training of specific tasks. To support multimodal input processing and fusion, the system relies on the pre-trained model library provided by Hugging Face, particularly models like BERT and CLIP for multimodal deep learning. Data preprocessing and augmentation are handled using open-source libraries such as OpenCV and librosa, with NumPy and Pandas used for data processing and management. During development, Docker 4.28 is employed for containerized deployment, ensuring environment portability and scalability. With this careful configuration of both hardware and software, the system can efficiently process large-scale data and perform complex deep learning tasks, ensuring stability and high efficiency in real-world applications.

4.2. Hyperparameters

In this experiment, the dataset is split into training, validation, and test sets with a standard ratio, where 70% of the data is used for training, 15% for validation, and 15% for testing. This division ensures that sufficient training data are available for model learning while the validation and test sets effectively assess the model's generalization ability. Regarding hyperparameter settings, several key hyperparameters of the model were adjusted and optimized to ensure optimal performance across different tasks. The learning rate α is set to 0.0001, selected through a grid search method from several candidate values. The batch size $batch_size$ is set to 32 to balance training efficiency and model performance. The Adam optimizer is used for training the deep learning models, with β_1 and β_2 set to 0.9 and 0.999, respectively, to ensure stable gradient updates. During the model training process, a 5-fold cross-validation strategy is employed to train and validate the model multiple times, avoiding overfitting and improving the model's stability and reliability. The results of cross-validation help further adjust the model's hyperparameters, optimizing its performance across different datasets. Through this rigorous hyperparameter tuning and cross-validation process, the system's performance in practical tasks is maximized.

4.3. Evaluation Metrics

In the evaluation of question-answering accuracy, common metrics include BLEU, ROUGE, Precision, Recall, and Accuracy, which help measure the performance of the system on a legal question set. First, BLEU is used to measure the n-gram matching

between the generated answer and the reference answer, while ROUGE evaluates the recall between the generated text and the reference text. Precision calculates the proportion of correct answers generated by the system, Recall measures the proportion of correct answers that the system can identify, and Accuracy provides an overall measure of the correctness of all predictions. Their mathematical formulas are as follows:

$$\text{BLEU}(n) = \min\left(1, \frac{p_n(\mathbf{S})}{\max(p_n(\mathbf{C}))}\right) \tag{24}$$

$$\text{ROUGE-L} = \frac{LCS(\mathbf{S}, \mathbf{C})}{\text{len}(\mathbf{C})} \tag{25}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{26}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{27}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{28}$$

where TP represents the number of correct answers identified by the system, FP represents the number of incorrectly identified answers, FN represents the number of missed correct answers, TN represents the number of correctly identified non-answers, $\text{LCS}(\mathbf{S}, \mathbf{C})$ represents the longest common subsequence length between the generated answer $\mathbf{S}$ and the reference answer $\mathbf{C}$, and $\text{len}(\mathbf{C})$ is the length of the reference answer.

In addition to system-level accuracy, three user-centered metrics are introduced to evaluate learning improvement and subjective experience: Understanding Improvement, Expression Improvement, and Satisfaction Score. These metrics are designed to quantify the system's impact on users' comprehension, expressive ability, and satisfaction in the context of legal English learning. Understanding Improvement measures the learner's progress in understanding legal scenarios and factual content before and after using the system. Expression Improvement reflects the extent to which learners enhance their legal English output, including grammatical correctness, use of terminology, and logical coherence. Satisfaction Score captures users' perceived experience with the system in terms of usability, feedback quality, and instructional value. The corresponding formulas for these metrics are defined as follows:

$$\text{Understanding Improvement} = \frac{S_{\text{post}} - S_{\text{pre}}}{S_{\text{pre}}} \tag{29}$$

$$\text{Expression Improvement} = \frac{E_{\text{post}} - E_{\text{pre}}}{E_{\text{pre}}} \tag{30}$$

$$\text{Satisfaction Score} = \frac{1}{N}\sum_{i=1}^{N} s_i \tag{31}$$

where S_{pre} and S_{post} represent the learner's average pre-test and post-test scores on case comprehension tasks, respectively. E_{pre} and E_{post} denote the average scores for expressive legal tasks before and after system interaction. s_i is the subjective satisfaction rating provided by the i-th participant on a five-point Likert scale and N is the total number of participants.

4.4. Baseline

In this experiment, the baseline models include three commonly used systems: the traditional ASR+SVM classifier [52], unimodal BERT question-answering system [53],

and image-retrieval-based question-answering model [54]. These models are meticulously selected to represent the typical paradigms of unimodal or weakly cross-modal question-answering systems. Specifically, the ASR+SVM classifier reflects a traditional pipeline approach based on speech-driven processing, while the unimodal BERT QA system exemplifies text-driven semantic reasoning. The image-retrieval-based QA model, on the other hand, serves as a representative of vision-driven query methods. These baselines were chosen to provide a comparative framework for evaluating the proposed method's performance gains in multimodal semantic alignment, deep cross-modal reasoning, and interactive feedback. By benchmarking against these classical systems, it becomes possible to clearly demonstrate the practical advantages of incorporating advanced multimodal fusion and cross-modal attention mechanisms. The observed improvements thus not only validate the effectiveness of the proposed framework but also highlight the limitations of shallow or unimodal strategies in complex legal education scenarios that require nuanced semantic understanding and evidentiary reasoning across modalities.

4.5. Experimental Results of Question-Answering Models

This experiment was designed to evaluate the overall performance of different question-answering models in the context of cross-modal legal English tasks, aiming to verify the system's capability to generate accurate responses after understanding complex case images and textual questions. The objective of the experiment was to compare the proposed method with traditional baseline methods, thereby validating the effectiveness of introducing multimodal fusion, visual-guided mechanisms, and deep semantic alignment strategies in improving answer accuracy, text generation quality, and overall reasoning ability. Five evaluation metrics, namely BLEU, ROUGE, Precision, Recall, and Accuracy, were utilized to comprehensively assess the performance, covering content generation quality, retrieval coverage, and classification accuracy, thereby fully reflecting the system's practical application potential in authentic legal education scenarios involving complex multimodal inputs.

As shown in Table 3, the Traditional ASR+SVM Classifier exhibited the weakest performance, primarily due to its separation of speech recognition and question–answer judgment processes, which failed to establish deep semantic connections between images and texts, resulting in lower BLEU and ROUGE scores. The image-retrieval-based QA model showed moderate improvement; however, its reliance on direct feature matching without text generation capability led to limitations in Precision and Recall. The unimodal BERT QA system achieved better generation quality through deep textual understanding, yet its lack of visual evidence incorporation resulted in incomplete comprehension when facing cross-modal problems. In contrast, the proposed method employed a three-modality deep encoding of vision, language, and speech and introduced a cross-modal attention mechanism to achieve high-level alignment among case images, question texts, and user responses, thus obtaining the best performance across all evaluation metrics. Mathematically, by utilizing contrastive learning and consistency optimization strategies, the proposed method enabled different modalities to form a more compact distribution within a unified feature space, significantly improving the modeling and reasoning ability for complex semantic structures.

Table 3. Experimental results of question-answering models.

Model	BLEU	ROUGE	Precision	Recall	Accuracy
Traditional ASR+SVM Classifier [52]	0.62	0.65	0.64	0.61	0.63
Image-Retrieval-based QA Model [54]	0.68	0.70	0.69	0.67	0.68
Unimodal BERT QA System [53]	0.74	0.77	0.76	0.74	0.75
Proposed Method	0.83	0.86	0.88	0.85	0.87

4.6. Experimental Results of Multimodal Matching Models

This experiment aimed to evaluate the performance of different cross-modal models in the image–text alignment task, validating the effectiveness of the proposed unified encoding framework that fuses visual evidence, textual questions, and spoken responses. The objective of the experiment was to compare the performance of mainstream cross-modal baseline models, including VisualBERT, CLIP, World2Vec, SpeechBERT, LXMERT, ALBEF, and BLIP, analyzing their strengths and weaknesses in multimodal feature extraction, alignment modeling, and retrieval reasoning, thereby demonstrating the proposed method's adaptability and improvement in handling complex legal scene alignments. Evaluation metrics included Matching Accuracy, Recall@1, Recall@5, and Mean Reciprocal Rank (MRR), comprehensively reflecting matching ability and reasoning depth from four perspectives: overall pairing accuracy, top-1 retrieval hit rate, multi-result retrieval coverage, and average ranking quality.

As shown in Table 4, VisualBERT exhibited the weakest performance across all metrics, mainly because its early-stage feature concatenation approach failed to effectively model fine-grained alignments between image regions and text segments. LXMERT introduced a separated encoder and cross-modal interaction mechanism, achieving slight improvement over VisualBERT; however, its single-step feature fusion strategy remained inadequate for complex image–text reasoning scenarios. Word2Vec showed marginal improvements in overall accuracy compared to VisualBERT, but due to its reliance on static, context-independent word embeddings, it lacked the representational capacity to capture semantic nuance in legal discourse or align with high-level visual concepts, resulting in limited Recall and MRR. SpeechBERT, leveraging speech–text joint modeling, achieved better cross-modal alignment by integrating BERT's semantic reasoning capabilities with audio inputs. However, the architecture's heavy dependence on speech signal quality and lack of visual grounding limited its robustness in tasks requiring multi-source evidentiary reasoning. CLIP demonstrated notable performance gains by leveraging contrastive learning with large-scale pretraining, thereby constructing a robust vision-language embedding space. Nonetheless, due to the absence of targeted fine-tuning, its effectiveness in handling nuanced legal evidence was limited, particularly in low-resource, high-reasoning scenarios. ALBEF further enhanced cross-modal representation by employing a vision–language pretraining strategy that integrated image–text matching, image–text contrastive learning, and masked language modeling in a unified framework. This design enabled better alignment and contextual understanding between modalities. However, ALBEF still relied on pre-extracted image features, which constrained its adaptability to domain-specific visual semantics in legal scenes. BLIP, as a more recent advancement, combined vision–language pretraining with a captioning-guided bootstrapping strategy, facilitating end-to-end learning of both unimodal and multimodal representations. Its architecture allowed for stronger text generation capabilities and more context-aware alignment. Consequently, BLIP outperformed previous baselines across all metrics and demonstrated robust reasoning ability in multimodal scenarios. The proposed method, through the fusion modeling of vision, text, and speech modalities and the introduction of dynamic attention and multi-level alignment optimization strategies, achieved deeper unification from the feature level to the semantic level. Different modalities were more compactly and discriminatively embedded in the feature space, leading to the best performance in Matching Accuracy, Recall@1, Recall@5, and MRR metrics. These results validated the effectiveness and superiority of the proposed mathematical mechanisms in cross-modal legal English question-answering scenarios.

Table 4. Experimental results of multimodal matching models.

Model	Matching Accuracy	Recall@1	Recall@5	MRR
VisualBERT	0.71	0.69	0.77	0.73
CLIP	0.79	0.77	0.84	0.80
Word2Vec	0.72	0.71	0.78	0.75
SpeechBERT	0.77	0.75	0.82	0.80
LXMERT	0.74	0.72	0.80	0.76
ALBEF	0.80	0.78	0.84	0.82
BLIP	0.81	0.80	0.86	0.83
Proposed Method	0.85	0.83	0.90	0.87

4.7. Experimental Results of User Study on Learning Improvement

A total of $N = 60$ participants were recruited for the user study, consisting of senior undergraduate and graduate students enrolled in legal English courses at three universities. Based on standardized English placement test results, participants were balanced by gender and language proficiency and then randomly assigned to one of four instructional groups, corresponding to the following teaching modes: traditional textbook-based instruction, conventional oral practice systems, a unimodal question-answering system, and the proposed multimodal question-answering system. Each group included 15 participants. Data collection was conducted over a four-week instructional period under a unified teaching schedule. Structured comprehension and expression assessments were administered both before and after instruction to measure learning progress. All participants completed case-based legal expression tasks of equivalent quantity and difficulty under instructor supervision. Scoring was performed independently by two legal education experts following the Common European Framework of Reference for Languages (CEFR) standards. Inter-rater reliability was validated using Cohen's $\kappa = 0.87$, indicating a high level of agreement. Learning outcomes were quantitatively evaluated across three dimensions: improvement in comprehension, improvement in legal expression, and user satisfaction. The first two dimensions were calculated as the difference between pre-test and post-test performance, while satisfaction scores were collected using a standardized five-point Likert scale. In addition, participants' open-ended feedback was systematically coded and analyzed using thematic analysis to identify recurring concerns regarding system functionality, interactivity, and perceived instructional value. For statistical analysis, repeated measures analysis of variance (ANOVA) was employed to assess both within-group learning gains and between-group differences. By systematically comparing the proposed method against traditional textbook instruction, oral practice systems, and unimodal QA models, the experiment comprehensively demonstrated the effectiveness of the proposed approach in enhancing multimodal reasoning capabilities, improving legal expression accuracy, and fostering learner motivation. This study was conducted through authentic classroom interventions to evaluate the impact of different instructional methods on learners' comprehension and expression gains, with the aim of validating the practical applicability and pedagogical value of the proposed cross-modal legal English question-answering system.

As shown in Table 5, the traditional textbook-based teaching method exhibited the weakest performance across all metrics, indicating that relying solely on textual materials was insufficient to stimulate learners' deep understanding of complex legal scenarios. Although the traditional oral practice system showed slight improvement in expression ability, the absence of multimodal evidence limited the overall enhancement of understanding. The unimodal question-answering system achieved moderate improvements in both understanding and expression by incorporating deep text comprehension mechanisms; however, the lack of visual evidence still resulted in an incomplete grasp of semantics.

In contrast, the proposed multimodal question-answering system integrated visual, linguistic, and auditory information into a unified semantic space, guiding learners to reason and express based on multimodal evidence, thereby achieving significant advantages across understanding, expression, and satisfaction metrics. Mathematically, multimodal feature alignment and dynamic feedback mechanisms effectively enhanced learners' cognitive load management and semantic transfer capabilities under complex input conditions, ensuring the system's high applicability and robust performance in real educational environments.

Table 5. Experimental results of user study on learning improvement.

Method	Understanding Improvement	Expression Improvement	Satisfaction Score
Textbook-based Teaching	0.45	0.40	0.60
Traditional Oral Practice System	0.52	0.48	0.65
Unimodal QA System	0.63	0.58	0.72
Proposed Method	0.78	0.75	0.88

4.8. Generalization Evaluation on Open Multimodal Dataset

To further verify the generalizability and robustness of the proposed method, we conducted additional experiments on the publicly available MultimodalQA dataset. This dataset features complex, multimodal legal question-answering tasks involving textual descriptions, visual references, and spoken queries. As presented in Table 6, the proposed method significantly outperformed baseline approaches across all five evaluation metrics: BLEU, ROUGE, Precision, Recall, and Accuracy.

Table 6. Performance on Open Dataset Multimodal QA.

Model	BLEU	ROUGE	Precision	Recall	Accuracy
Traditional ASR+SVM	0.62	0.64	0.63	0.60	0.63
Image Retrieval QA	0.67	0.68	0.68	0.65	0.67
Unimodal BERT QA	0.73	0.76	0.75	0.73	0.73
Proposed Method	**0.83**	**0.85**	**0.87**	**0.83**	**0.86**

Traditional ASR+SVM models demonstrated limited performance due to their modular pipeline and lack of deep semantic alignment. The image retrieval QA baseline showed slight improvements by incorporating visual cues but remained constrained by shallow retrieval logic. The unimodal BERT QA model achieved better performance by leveraging pretrained language understanding, yet it failed to effectively integrate cross-modal context. In contrast, the proposed method, which jointly encodes and aligns textual, visual, and auditory modalities within a unified framework, achieved superior performance across all metrics (BLEU: 0.83, ROUGE: 0.85, Precision: 0.87, Recall: 0.83, Accuracy: 0.86). These results demonstrate the method's strong capacity for multimodal reasoning and reinforce its adaptability in complex, open-domain legal QA scenarios beyond the in-house dataset.

4.9. Ablation Study on Learners with Varying English Proficiency

To further examine the adaptability and effectiveness of the proposed system across user groups with diverse language abilities, we conducted a user study stratified by English proficiency levels. Participants were categorized into three groups, high-level, medium-level, and low-level learners, based on standardized language assessment scores. The evaluation focused on three key indicators: understanding improvement, expression improvement, and user satisfaction, as shown in Table 7.

Table 7. User study on learners with different English proficiency levels.

Group	Understanding Improvement	Expression Improvement	Satisfaction Score
High-Level Learners	0.81	0.79	0.89
Medium-Level Learners	0.76	0.74	0.86
Low-Level Learners	0.72	0.69	0.83

The results demonstrate that all groups benefited from the system, with noticeable improvements in both comprehension and expressive skills. High-level learners showed the highest gains (understanding: 0.81, expression: 0.79), reflecting their greater ability to integrate multimodal cues and utilize feedback for refinement. Medium-level learners also exhibited consistent improvements (0.76 and 0.74), indicating that the system supports intermediate learners in bridging semantic gaps during multimodal interactions. Although the low-level group achieved relatively lower gains (0.72 and 0.69), their satisfaction score remained high (0.83), suggesting that the system's interface and scaffolding mechanisms provided accessible and encouraging learning experiences, even for users with limited language proficiency. These findings indicate that the proposed method maintains stable usability and pedagogical effectiveness across a wide range of learner profiles, supporting its broader applicability in real-world legal English education settings.

4.10. Discussion

4.10.1. Practical Implications and Performance Analysis

A cross-modal legal English question-answering system based on visual and acoustic sensor inputs was proposed to address the limitations of traditional legal English teaching, such as reliance on single-modality input, insufficient interactivity, and difficulties in semantic transfer. Through systematic experimental designs and multidimensional evaluation metrics, the proposed method was validated for its significant advantages in complex case comprehension, legal expression training, and enhancement of learning experiences. In the evaluation of question-answering accuracy, traditional ASR+SVM classifiers and retrieval-based QA models exhibited evident limitations when handling cross-modal inputs, particularly due to their lack of deep semantic reasoning and visual evidence modeling capabilities. Although the unimodal BERT system demonstrated improvements in textual understanding, its neglect of image and audio inputs resulted in insufficient reasoning depth. In contrast, the proposed method, through a fusion mechanism of vision, language, and speech modalities and the introduction of a cross-modal attention mechanism, significantly enhanced the system's capability to capture complex relationships between case evidence and legal questions, thus achieving superior performance across BLEU, ROUGE, Precision, Recall, and Accuracy metrics. In the analysis of multimodal matching performance, the comparison among different models further highlighted the critical role of cross-modal fusion strategies in improving image–text alignment. Although VisualBERT and LXMERT initially achieved semantic mapping between images and text, their lack of fine-grained alignment and optimization led to notable deficiencies in matching accuracy and retrieval performance. CLIP, while establishing a stronger unified cross-modal space through large-scale contrastive learning, lacked fine-tuning for small-sample complex legal reasoning tasks, leading to performance losses on fine-grained evidence tasks. By introducing dynamic attention mechanisms and a unified encoding space design, the proposed method reinforced deep fine-grained cross-modal relationship modeling, effectively addressing challenges such as fragmented evidence and long reasoning chains in legal image–text scenarios. Consequently, the proposed method achieved the best performance across Matching Accuracy, Recall@1, Recall@5, and MRR metrics, demon-

strating that shallow feature fusion is insufficient for understanding complex scenarios and deep multimodal fusion with structured reasoning is indispensable.

The results of the user study further confirmed that improvements in model performance translated into tangible enhancements in educational effectiveness. In real classroom environments, the proposed method significantly enhanced learners' case comprehension depth and English legal expression accuracy, with satisfaction scores markedly surpassing those of traditional and unimodal teaching methods. Traditional textbook- or oral-based methods, due to a lack of contextual sensing and interactive reasoning, showed limited learning effectiveness, while unimodal QA systems improved some aspects of expression but failed to cultivate the systematic reasoning abilities necessary for complex evidence-based reasoning. Through unified reasoning supported by visual, linguistic, and auditory modalities, learners naturally constructed factual narratives and legal reasoning chains, demonstrating superior language transfer and critical thinking capabilities. These results reflect the practical value and application potential of intelligent multimodal perception and deep fusion technologies in legal English education, offering empirical validation and future references for the design of intelligent multimodal interactive systems in foreign language education.

4.10.2. Methodological Reflections and Scientific Contributions

While the practical advantages of the proposed system have been substantiated through quantitative metrics and user studies, it is equally important to reflect on its methodological and scientific contributions within the broader context of multimodal research.

Recent research has increasingly emphasized the theoretical and empirical advantages of multimodal architectures over unimodal or bimodal counterparts. Multimodal learning enables richer semantic representation by leveraging complementary information across modalities, enhancing generalization in both perception and reasoning tasks [55]. For instance, studies in visual question answering and grounded language understanding have shown that combining vision and language modalities improves spatial inference and referential accuracy [56]. Similarly, integrating acoustic information has been found to reinforce temporal modeling and improve robustness to ambiguous textual or visual cues [57]. In educational contexts, multimodal approaches have demonstrated improvements in learner engagement, feedback personalization, and cognitive load management, particularly when integrated with adaptive feedback systems [58]. These findings suggest that multimodal systems are not merely additive in capability but fundamentally more expressive and cognitively aligned with human information processing.

This work integrates three distinct modalities—vision, language, and speech—into a unified representation space, extending prior efforts that typically focused on bimodal fusion [19,20]. The use of dynamic attention mechanisms, affordance-driven visual masking, and multi-level alignment losses contributes to a deeper, fine-grained modeling of heterogeneous legal inputs. These strategies address key challenges in multimodal learning, such as modality imbalance, semantic drift, and evidence grounding, which are often overlooked in earlier designs such as the VisualBERT [24] or wav2vec-only pipelines [25]. Furthermore, unlike traditional question-answering systems that adopt shallow concatenation or single-stage fusion, the proposed framework introduces structured semantic constraints (e.g., contrastive alignment loss, consistency fusion loss), enabling hierarchical alignment from the feature to semantic levels. This aligns with recent scientific perspectives emphasizing the need for "structured explainability" and transparent decision pathways in multimodal educational systems [18,21]. Compared to baseline models, the inclusion of a speech-informed semantic parsing module in conjunction with image-based evidence understanding represents a novel direction. Prior studies have separately validated the

role of vision–text [59,60] and speech–text models [61], but the legal domain presents a unique challenge where factual accuracy, logical rigor, and multimodal alignment must co-exist—an area underexplored in the current research. In this context, the proposed integration strategy enhances the interpretability, adaptability, and semantic depth required for high-stakes, domain-specific applications.

In sum, this study bridges a critical gap in the literature by introducing a scientifically grounded, pedagogically oriented multimodal learning system tailored for legal education. It demonstrates that educational AI models benefit not only from larger datasets or more parameters, but from principled integration of cross-modal structure, interpretability mechanisms, and domain-aware representation strategies.

4.11. Limitations and Future Work

Despite the strong performance demonstrated by the proposed cross-modal legal English question-answering system based on visual and auditory sensor inputs in terms of question-answering accuracy, multimodal matching capability, and user learning outcomes, several limitations remain. At present, the system primarily focuses on three data types: case images, textual questions, and spoken responses. More complex legal document forms, such as judicial opinions, contractual clauses, and courtroom transcripts, have not yet been incorporated, which limits the model's applicability in tasks with high semantic complexity. In addition, the current system remains primarily designed for English-language contexts, and its transferability and generalization ability in multilingual settings require further investigation. Although the principles of informed consent and anonymization were followed during data collection, ethical and privacy risks associated with image and speech sensor inputs remain a critical concern in real-world educational deployments. Future iterations of the system will integrate data desensitization mechanisms, differential privacy algorithms, and ethical review procedures to ensure compliance, transparency, and user controllability. In terms of functional expansion, subsequent research will further explore model compression and inference optimization strategies suitable for deployment on resource-constrained devices. Additional sensor modalities, such as tactile feedback, eye-tracking, and physiological signals, will be introduced to construct a more immersive and realistic language learning experience. On this basis, future studies will focus on the architectural exploration of next-generation cross-modal frameworks, including the integration of graph neural networks with multi-scale attention mechanisms, the construction of multimodal causal reasoning models, and the implementation of dynamic context modeling strategies. These methods are expected to enhance the model's capacity to capture structured semantic relationships across heterogeneous data sources and to advance multimodal intelligent language learning systems toward stronger generalization and interactivity, thereby extending the theoretical boundaries of human–AI collaborative language education.

5. Conclusions

With the rapid development of artificial intelligence and sensor technologies, traditional legal English teaching models have shown increasing limitations in multimodal understanding, interactive reasoning, and language transfer capabilities. To address the challenges of relying solely on text-based input, a lack of semantic grounding in perceptual evidence, and falling short in real-world legal scenarios, this paper proposed a cross-modal legal English question-answering system based on visual and acoustic sensor inputs. The core innovation lies in the construction of a unified vision–language–speech fusion framework, enhanced by dynamic attention mechanisms and multimodal consistency optimization strategies. This design enables deep perception of complex legal cases

and supports structured training in legal reasoning and English expression. Extensive experiments demonstrate that the proposed system significantly outperforms traditional baselines—such as ASR+SVM classifiers, image-retrieval-based models, and unimodal BERT systems—achieving an Accuracy of 0.87, a Precision of 0.88, and a Recall of 0.85. User studies further confirm its educational effectiveness, with learners' comprehension, expressive ability, and satisfaction scores significantly improved in real teaching scenarios. From a theoretical perspective, this study contributes to advancing multimodal semantic learning in legal language education by integrating structured perception, language alignment, and feedback generation within an interactive learning framework. Moreover, the incorporation of semantic feedback and image–speech alignment mechanisms provides a path toward improving system interpretability and learner trust. Nonetheless, certain limitations remain. The current system focuses on a limited set of input modalities and domain-specific tasks, and its generalizability to more complex legal texts, multilingual environments, and cross-cultural scenarios remains to be validated. In addition, ethical and privacy issues surrounding sensor-based data collection call for stricter compliance frameworks in real deployments. Future research will explore lightweight model architectures suitable for resource-constrained environments, incorporate additional sensor modalities (e.g., haptic and gaze data), and integrate privacy-preserving mechanisms such as data anonymization and differential privacy. More importantly, new research directions will focus on the development of adaptable multimodal architectures, including graph-based reasoning networks, causal multimodal models, and dynamic context-aware learning frameworks, thereby pushing the boundaries of intelligent language education systems. In summary, this work not only presents a practical and effective solution to current challenges in legal English learning but also provides a robust theoretical and technical foundation for future advances in multimodal intelligent education systems.

Author Contributions: Conceptualization, Y.C., C.H., S.G., D.B. and C.L.; Data curation, Y.L.; Formal analysis, X.C. and S.L.; Funding acquisition, C.L.; Investigation, S.L.; Methodology, Y.C., C.H. and S.G.; Project administration, D.B. and C.L.; Resources, Y.L.; Software, Y.C., C.H., S.G. and X.C.; Supervision, D.B. and C.L.; Validation, X.C. and S.L.; Visualization, Y.L.; Writing—original draft, Y.C., C.H., S.G., Y.L., X.C., S.L., D.B. and C.L.; Y.C., C.H. and S.G. contributed equally to this work. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China, grant number 61202479.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available upon request from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Grifoni, P.; D'ulizia, A.; Ferri, F. When language evolution meets multimodality: Current status and challenges toward multimodal computational models. *IEEE Access* **2021**, *9*, 35196–35206. [CrossRef]
2. Garner, B.A. *Legal Writing in Plain English: A Text with Exercises*; University of Chicago Press: Chicago, IL, USA, 2023.
3. Chalkidis, I.; Jana, A.; Hartung, D.; Bommarito, M.; Androutsopoulos, I.; Katz, D.M.; Aletras, N. LexGLUE: A benchmark dataset for legal language understanding in English. *arXiv* **2021**, arXiv:2110.00976. [CrossRef]
4. Trott, K.C. Examining the Effects of Legal Articulation in Memory Accuracy. Master's Thesis, The University of Regina (Canada), Regina, SK, Cananda, 2022.
5. Danylchenko-Cherniak, O. Applying Disign Thinking Principles to Modern Legal English Education. *Philol. Treatises* **2024**, *16*, 32–50. [CrossRef]

6. Pasa, B.; Sinni, G. New Frontiers of Legal Knowledge: How Design Provotypes Can Contribute to Legal Change. Ducato, R., Strowel, A., Marique E., Eds.; 2024. p. 27. Available online: https://www.torrossa.com/en/resources/an/5826118#page=27 (accessed on 25 May 2025).
7. Sheredekina, O.; Karpovich, I.; Voronova, L.; Krepkaia, T. Case technology in teaching professional foreign communication to law students: Comparative analysis of distance and face-to-face learning. *Educ. Sci.* **2022**, *12*, 645. [CrossRef]
8. Salmerón-Manzano, E.; García, A.A.; Zapata, A.; Manzano-Agugliaro, F. Generic Skills Assessment in The Master Of Laws At University Of Almeria. In Proceedings of the INTED2021 Proceedings, Online, 8–9 March 2021; IATED: Valencia, Spain, 2021; pp. 1154–1163.
9. Mazer, L.; Varban, O.; Montgomery, J.R.; Awad, M.M.; Schulman, A. Video is better: Why aren't we using it? A mixed-methods study of the barriers to routine procedural video recording and case review. *Surg. Endosc.* **2022**, *36*, 1090–1097. [CrossRef] [PubMed]
10. Hloviuk, I.; Zavtur, V.; Zinkovskyy, I.; Fedorov, O. The use of video and audio recordings provided by victims of domestic violence as evidence. *Soc. Leg. Stud.* **2024**, *1*, 145–154. [CrossRef]
11. Mangaroska, K.; Martinez-Maldonado, R.; Vesin, B.; Gašević, D. Challenges and opportunities of multimodal data in human learning: The computer science students' perspective. *J. Comput. Assist. Learn.* **2021**, *37*, 1030–1047. [CrossRef]
12. Lewis, K.O.; Popov, V.; Fatima, S.S. From static web to metaverse: Reinventing medical education in the post-pandemic era. *Ann. Med.* **2024**, *56*, 2305694. [CrossRef]
13. Al-Tarawneh, A.; Al-Badawi, M.; Hatab, W.A. Professionalizing legal translator training: Prospects and opportunities. *Theory Pract. Lang. Stud.* **2024**, *14*, 541–549. [CrossRef]
14. Nhac, T.H. Enhancing Legal English skills for Law students through simulation-based activities. *Int. J. Learn. Teach. Educ. Res.* **2023**, *22*, 533–549. [CrossRef]
15. Soto, J.H.B. Technology-Enhanced Approaches for Teaching English to Law Students: Innovations and Best Practices. *Cienc. Lat. Rev. Multidiscip.* **2024**, *8*, 4567–4592.
16. Xatamova, N.; Ashurov, J. The Future of Legal English Learning: Integrating AI into ESP Education. *SPAST Rep.* **2024**, *1*, 87–102.
17. Peng, Y.T.; Lei, C.L. Using Bidirectional Encoder Representations from Transformers (BERT) to predict criminal charges and sentences from Taiwanese court judgments. *PeerJ Comput. Sci.* **2024**, *10*, e1841. [CrossRef] [PubMed]
18. Schmidt, T.; Strasser, T. Artificial intelligence in foreign language learning and teaching: A CALL for intelligent practice. *Angl. Int. J. Engl. Stud.* **2022**, *33*, 165–184.
19. Tang, H.; Hu, Y.; Wang, Y.; Zhang, S.; Xu, M.; Zhu, J.; Zheng, Q. Listen as you wish: Fusion of audio and text for cross-modal event detection in smart cities. *Inf. Fusion* **2024**, *110*, 102460. [CrossRef]
20. Yang, X.; Tan, L. Multimodal knowledge graph construction for intelligent question answering systems: Integrating text, image, and audio data. *Aust. J. Electr. Electron. Eng.* **2025**, 1–10. [CrossRef]
21. Al-Saadawi, H.F.T.; Das, B.; Das, R. A systematic review of trimodal affective computing approaches: Text, audio, and visual integration in emotion recognition and sentiment analysis. *Expert Syst. Appl.* **2024**, *255*, 124852. [CrossRef]
22. Kim, D.; Kang, P. Cross-modal distillation with audio–text fusion for fine-grained emotion classification using BERT and Wav2vec 2.0. *Neurocomputing* **2022**, *506*, 168–183. [CrossRef]
23. Yun, L. Application of English semantic understanding in multimodal machine learning. In Proceedings of the 2024 International Conference on Artificial Intelligence, Deep Learning and Neural Networks (AIDLNN), Guangzhou, China, 22–24 September 2024; pp. 233–237.
24. Bandyopadhyay, D.; Hasanuzzaman, M.; Ekbal, A. Seeing through VisualBERT: A causal adventure on memetic landscapes. *arXiv* **2024**, arXiv:2410.13488.
25. Lim, J.; Kim, K. Wav2vec-vc: Voice conversion via hidden representations of wav2vec 2.0. In Proceedings of the ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 14–19 April 2024; pp. 10326–10330.
26. Bonaccorsi, D. Speech-Text Cross-Modal Learning through Self-Attention Mechanisms. Ph.D. Thesis, Politecnico di Torino, Torino, Italy, 2023.
27. Wang, Y.; Xiao, B.; Bouferguene, A.; Al-Hussein, M.; Li, H. Vision-based method for semantic information extraction in construction by integrating deep learning object detection and image captioning. *Adv. Eng. Inform.* **2022**, *53*, 101699. [CrossRef]
28. Ge, Y.; Ge, Y.; Liu, X.; Wang, J.; Wu, J.; Shan, Y.; Qie, X.; Luo, P. Miles: Visual bert pre-training with injected language semantics for video-text retrieval. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 691–708.
29. Nguyen, C.D.; Vu-Le, T.A.; Nguyen, T.; Quan, T.; Luu, A.T. Expand BERT representation with visual information via grounded language learning with multimodal partial alignment. In Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, ON, Canada, 29 October–3 November 2023; pp. 5665–5673.

30. Li, Y.; Liang, F.; Zhao, L.; Cui, Y.; Ouyang, W.; Shao, J.; Yu, F.; Yan, J. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv* **2021**, arXiv:2110.05208.
31. Lee, J.; Kim, J.; Shon, H.; Kim, B.; Kim, S.H.; Lee, H.; Kim, J. Uniclip: Unified framework for contrastive language-image pre-training. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 1008–1019.
32. Pan, X.; Ye, T.; Han, D.; Song, S.; Huang, G. Contrastive language-image pre-training with knowledge graphs. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 22895–22910.
33. Pepino, L.; Riera, P.; Ferrer, L. Emotion recognition from speech using wav2vec 2.0 embeddings. *arXiv* **2021**, arXiv:2104.03502.
34. Cai, J.; Song, Y.; Wu, J.; Chen, X. Voice disorder classification using Wav2vec 2.0 feature extraction. *J. Voice* **2024**, *in press*. [CrossRef]
35. Sadhu, S.; He, D.; Huang, C.W.; Mallidi, S.H.; Wu, M.; Rastrow, A.; Stolcke, A.; Droppo, J.; Maas, R. Wav2vec-c: A self-supervised model for speech representation learning. *arXiv* **2021**, arXiv:2103.08393.
36. Chen, L.; Deng, Y.; Wang, X.; Soong, F.K.; He, L. Speech bert embedding for improving prosody in neural tts. In Proceedings of the ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 6563–6567.
37. Jiang, Y.; Sharma, B.; Madhavi, M.; Li, H. Knowledge distillation from bert transformer to speech transformer for intent classification. *arXiv* **2021**, arXiv:2108.02598.
38. Xie, J.; Zhao, Y.; Zhu, D.; Yan, J.; Li, J.; Qiao, M.; He, G.; Deng, S. A machine learning-combined flexible sensor for tactile detection and voice recognition. *ACS Appl. Mater. Interfaces* **2023**, *15*, 12551–12559. [CrossRef]
39. Rita, P.; Ramos, R.; Borges-Tiago, M.T.; Rodrigues, D. Impact of the rating system on sentiment and tone of voice: A Booking. com and TripAdvisor comparison study. *Int. J. Hosp. Manag.* **2022**, *104*, 103245. [CrossRef]
40. Liu, Y.; Zhang, Y.A.; Zeng, M.; Zhao, J. A novel distance measure based on dynamic time warping to improve time series classification. *Inf. Sci.* **2024**, *656*, 119921. [CrossRef]
41. Yu, S.; Fang, C.; Yun, Y.; Feng, Y. Layout and image recognition driving cross-platform automated mobile testing. In Proceedings of the 2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE), Virtual, 25–28 May 2021; pp. 1561–1571.
42. al Khatib, A.; Shaban, M.S. Innovative Instruction of Law Teaching and Learning: Using Visual Art, Creative Methods, and Technology to Improve Learning. *Int. J. Membr. Sci. Technol.* **2023**, *10*, 361–371. [CrossRef]
43. Balayn, A.; Soilis, P.; Lofi, C.; Yang, J.; Bozzon, A. What do you mean? Interpreting image classification with crowdsourced concept extraction and analysis. In Proceedings of the Web Conference 2021, Ljubljana, Slovenia, 19–23 April 2021; pp. 1937–1948.
44. Zhang, J.; Hu, J. Enhancing English education with Natural Language Processing: Research and development of automated grammar checking, scoring systems, and dialogue systems. *Appl. Comput. Eng.* **2024**, *102*, 12–17. [CrossRef]
45. Nypadymka, A.; Danchenko, L. From motivation to mastery: English language learning in legal education. In Proceedings of the Science and Education in the Conditions of Today's Challenges, International Scientific and Practical Conference, Chernihiv, Ukraine, 23 May 2023; pp. 342–346.
46. Danylchenko-Cherniak, O. Digital approaches in legal English education: Bridging tradition and technology. *Stud. Comp. Educ.* **2024**, 77–83.
47. Martinez-Gil, J. A survey on legal question–answering systems. *Comput. Sci. Rev.* **2023**, *48*, 100552. [CrossRef]
48. Yang, X.; Wang, Z.; Wang, Q.; Wei, K.; Zhang, K.; Shi, J. Large language models for automated q&a involving legal documents: A survey on algorithms, frameworks and applications. *Int. J. Web Inf. Syst.* **2024**, *20*, 413–435.
49. Hockly, N. Artificial intelligence in English language teaching: The good, the bad and the ugly. *Relc J.* **2023**, *54*, 445–451. [CrossRef]
50. Moorhouse, B.L.; Li, Y.; Walsh, S. E-classroom interactional competencies: Mediating and assisting language learning during synchronous online lessons. *Relc J.* **2023**, *54*, 114–128. [CrossRef]
51. Bellei, C.; Muñoz, G. Models of regulation, education policies, and changes in the education system: A long-term analysis of the Chilean case. *J. Educ. Chang.* **2023**, *24*, 49–76. [CrossRef]
52. García-Cabellos, J.M.; Peláez-Moreno, C.; Gallardo-Antolín, A.; Pérez-Cruz, F.; Díaz-de María, F. SVM classifiers for ASR: A discussion about parameterization. In Proceedings of the 2004 12th European Signal Processing Conference, Vienna, Austria, 6–10 September 2004; pp. 2067–2070.
53. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; Volume 1, pp. 4171–4186.
54. Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C.L.; Parikh, D. Vqa: Visual question answering. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2425–2433.
55. Guo, W.; Wang, J.; Wang, S. Deep Multimodal Representation Learning: A Survey. *IEEE Access* **2019**, *7*, 63373–63394. [CrossRef]
56. Yang, S.; Luo, S.; Han, S.C.; Hovy, E. MAGIC-VQA: Multimodal And Grounded Inference with Commonsense Knowledge for Visual Question Answering. *arXiv* **2025**, arXiv:2503.18491.

57. Xing, Z.; Hu, X.; Fu, C.W.; Wang, W.; Dai, J.; Heng, P.A. EchoInk-R1: Exploring Audio-Visual Reasoning in Multimodal LLMs via Reinforcement Learning. *arXiv* **2025**, arXiv:2505.04623.
58. Bewersdorff, A.; Hartmann, C.; Hornberger, M.; Seßler, K.; Bannert, M.; Kasneci, E.; Kasneci, G.; Zhai, X.; Nerdel, C. Taking the next step with generative artificial intelligence: The transformative role of multimodal large language models in science education. *Learn. Individ. Differ.* **2025**, *118*, 102601. [CrossRef]
59. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models From Natural Language Supervision. *arXiv* **2021**, arXiv:2103.00020.
60. Li, J.; Selvaraju, R.R.; Gotmare, A.D.; Joty, S.; Xiong, C.; Hoi, S. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. *arXiv* **2021**, arXiv:2107.07651.
61. Chuang, Y.S.; Liu, C.L.; Lee, H.Y.; Lee, L.S. SpeechBERT: An Audio-and-text Jointly Learned Language Model for End-to-end Spoken Question Answering. *arXiv* **2020**, arXiv:1910.11559.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

sensors

MDPI

Article

Analysis of Pleasure and Displeasure in Harmony Between Colored Light and Fragrance by the Left and Right OFC Response Differences

Toshinori Oba [1], Midori Tanaka [2] and Takahiko Horiuchi [2,*]

[1] Graduate School of Science and Engineering, Chiba University, Chiba 263-8522, Japan
[2] Graduate School of Informatics, Chiba University, Chiba 263-8522, Japan
* Correspondence: horiuchi@faculty.chiba-u.jp

Highlights

What are the main findings?

- By sensing biological signals, we have experimentally revealed that the left OFC is activated when perceiving a harmonious combination of colored light and fragrance, whereas the right OFC is activated when the combination is disharmonious.
- By sensing biological signals, we have experimentally revealed that functional differences between the left and right OFC are influenced not only by the pleasure–displeasure axis but also by the crossmodal–multimodal perception axis.

What is the implication of the main finding?

- We can develop a scientific understanding of the emotional impact of visual and olfactory interactions based on biological information sensing.
- Our study lays the groundwork for further developments in sensor-based neuroscience and biomedical engineering.

Abstract: Daily actions are influenced by sensory information. Several studies have investigated the multisensory integration of multiple sensory modalities, known as crossmodal perception. Recently, visual–olfactory crossmodal perception has been studied using objective physiological measures rather than subjective evaluations. This study focused on sensing in the orbitofrontal cortex (OFC), which responds to visual and olfactory stimuli, and may serve as a physiological indicator of perception. Using near-infrared spectroscopy (NIRS), we analyzed the emotions evoked by combinations of colored light and fragrance with a particular focus on the lateralization of brain function. We selected pleasant and unpleasant fragrances from some essential oils, paired with colored lights that were perceived as either harmonious or disharmonious with the fragrances. NIRS measurements were conducted under the four following conditions: fragrance-only, colored light-only, harmonious crossmodal, and disharmonious crossmodal presentations. The results showed that the left OFC was activated during the crossmodal presentation of a harmonious color with a pleasant fragrance, thereby evoking pleasant emotions. In contrast, during the crossmodal presentation of a disharmonious color with an unpleasant fragrance, the right OFC was activated, suggesting increased displeasure. Additionally, the lateralization of brain function between the left and right OFC may be influenced by 'pleasure–displeasure ' and 'crossmodal perception–multimodal perception'.

Keywords: crossmodal perception; olfactory; color; harmony; emotion; orbitofrontal cortex; near-infrared spectroscopy

1. Introduction

We obtain various types of information every day using the five senses and select appropriate actions in response. For example, we can simultaneously perceive multiple sensory inputs such as 'video and sound' or 'color and scent'. We then integrate this information to select appropriate actions. While the five senses have been investigated in studies from various domains, the number of studies on multisensory integration, also referred to as crossmodal or multimodal integration, has increased in recent years [1]. Both crossmodal and multimodal are terms related to multiple modalities, but they differ in meaning and usage. Crossmodal generally refers to situations where multisensory integration occurs through interactions between different modalities. Here, "interaction" means that two or more perceptual stimuli influence each other. Specifically, this includes cases where a stimulus in one modality affects the perception or cognition of another modality, as well as instances where such influences occur bidirectionally. On the other hand, multimodal refers to situations where multiple modalities coexist and are integrated without necessarily interacting with each other. In other words, the key difference between crossmodal and multimodal integration lies in whether the modalities influence one another during integration or remain independent. This study analyzes crossmodal perception, where interactions between different modalities influence perception.

Traditional crossmodal study has primarily focused on vision [2]. However, in recent decades, crossmodal interactions involving other senses, such as olfaction, have also been discussed. Even within the scope of visual–olfactory crossmodal perception, numerous studies have been reported [3–17]. Zellner et al. [3] focused on the visual–olfactory crossmodal perception and investigated the effect of color on the perceived intensities of specific food scents. Their experiment involved participants evaluating the intensity of scents in a red liquid and a colorless liquid, each containing a strawberry scent. The results indicated a tendency for the colored liquid to be perceived as having a stronger scent than the colorless liquid. Gilbert et al. [4] confirmed significant color associations for 20 different fragrances that remained stable, even when the experiment was repeated years later, suggesting a strong correlation between vision and olfaction. Miura et al. [5] used color patches to investigate the effects of color–fragrance harmony on stress reduction. They found that harmonious color–fragrance combinations reduce stress more than disharmonious combinations. In addition, some studies have focused on the speeds of processing and discriminating scent. Demattè et al. [6] presented participants with scents and color patches and instructed them to discriminate each stimulus as quickly and accurately as possible. The stimuli were presented under either a compatible condition (e.g., a strawberry scent with pink or a spearmint scent with turquoise) or an incompatible condition (e.g., a strawberry scent with turquoise or a spearmint scent with pink). Reaction times and identification accuracy were measured. The results showed that the more the color and scent were matched, the faster and more accurate the responses became. In addition, Demattè et al. [7] showed that when the color and shape are congruent with a scent, the accuracy of scent discrimination increases, whereas incongruence leads to a decrease in accuracy. Various studies have demonstrated the existence of crossmodal associations between vision and olfaction.

While previous studies have relied primarily on subjective evaluations, recent studies have increasingly focused on objective analyses using physiological responses [8–17]. Österbauer et al. [8] investigated the neurophysiological correlates of visual–olfactory crossmodal perception using functional magnetic resonance imaging (fMRI). Their experiment showed that the orbitofrontal cortex (OFC) and the caudal region of the insular cortex exhibited greater activation as the congruence between scent and color pairs increased. Furthermore, numerous studies have measured physiological responses using near-infrared spectroscopy

(NIRS), which has a higher temporal resolution than fMRI [9–17]. Compared to existing neuroimaging techniques, NIRS is relatively inexpensive, easy to set up, and does not require a specialized environment. Its high motion tolerance allows studies to be conducted in natural settings. Pinti et al. [9] stated that NIRS allows the study of visual–olfactory crossmodal perception in natural environments, generating more reliable and ecologically valid data. NIRS measures brain activity at a maximum depth approximately halfway between the light source and detector [10]. The maximum distance between the light source and detector to maintain detectable signals was 3 cm, resulting in a measurement depth of approximately 1.5 cm from the scalp surface. Therefore, it is challenging to use NIRS to measure the deep brain regions associated with olfactory perception, such as the piriform cortex, insular cortex, and hippocampus [11]. However, among the regions associated with olfactory and visual perception, there are relatively shallow areas, such as the OFC, which have been investigated by recent studies using NIRS [12,13].

The connection between colored light and olfactory sensing is an emerging topic with important implications for sensor-based research and applications. Our study contributes to this area by demonstrating how NIRS can be utilized to detect the neural responses associated with multisensory integration. Understanding these interactions is crucial for the following reasons:

1. Advancements in Sensory Monitoring Technologies

Investigating how different sensory modalities interact provides essential insights for developing advanced sensor systems. For example, our findings could inform the design of multimodal sensing technologies used in healthcare, smart environments, and human-computer interaction.

2. Enhancing Sensor-Based Brain Activity Monitoring

By applying NIRS to study multisensory interactions, our work contributes to refining non-invasive brain sensing methodologies. These insights could support further developments in real-time brain monitoring, particularly in areas like neurofeedback, assistive technologies, and affective computing.

In previous studies on visual–olfactory crossmodal perception, various types of colors have been used as visual stimuli, such as colored liquids [3,14], color patches [4–7], and monitor displays [16], all of which are small-scale visual stimuli of a few centimeters by a few centimeters, focusing on flat visual stimuli. However, since we live in environments with ambient lighting daily, spatial testing is important. Yamashita et al. [16] and Oba et al. [17] used physiological indicators to analyze the color–fragrance harmony in lighting environments. Colored light was used as the color stimulus for spatial examination, and an objective analysis was performed by measuring the OFC response to the stimuli. Yamashita et al. [16] initially conducted a preliminary survey using five fragrances and 36 colored lights to select a 'pleasant fragrance' and 'unpleasant fragrance', as well as 'harmonious colors' and 'disharmonious colors' for each participant. The OFC response was measured under the four following conditions: (1) fragrance-only, (2) colored light-only, (3) harmonious crossmodal, and (4) disharmonious crossmodal presentations. The results showed that the highest OFC activation occurred during the fragrance-only presentation, with the left OFC showing greater activation than the right OFC. This suggested a relationship between pleasant emotions and the left OFC. Furthermore, Oba et al. [17] analyzed whether differences in the lighting environment affected OFC activation and whether the left and right OFC activated differently based on pleasure or displeasure. They found that the OFC was more activated in a relaxing lighting environment closer to real-life conditions. Moreover, pleasant emotions were associated with greater activation of the left OFC, which supported the results of Yamashita et al. [16]. However, these studies

used a limited number of fragrances as stimuli, and these were generally perceived as pleasant by the participants. Therefore, the analysis focused on pleasant emotions related to the left OFC, making it challenging to analyze unpleasant emotions. This study aimed to analyze the pleasure–displeasure emotions evoked by color–fragrance harmony while also considering unpleasant emotions based on physiological indicators and focusing on the lateralization of brain function.

Our work contributes to the sensing field in two key ways:

1. Application of NIRS in Multisensory Research

We demonstrated the effectiveness of NIRS as a non-invasive tool to measure cortical activity related to the interaction between light and olfactory stimuli. This application showcases the potential of NIRS in advancing real-time sensory monitoring, expanding its use beyond traditional domains.

2. Novelty in Sensing Approach

Although we did not develop a new sensor, our methodological approach integrates NIRS into a novel approach to explore sensory processing. Specifically, we applied NIRS to capture the subtle hemodynamic responses associated with crossmodal interactions, which could inspire further advancements in sensor-based neuroscience research.

2. Experiment I: Analysis of Pleasure and Displeasure in Harmony of Colored Light and Fragrance

2.1. Measurement of Orbitofrontal Cortex Responses

Recent psychophysical experiments dealing with emotions have increased the use of physiological indicators such as brain waves for objective analysis. In previous studies [9–17], the OFC was used as a physiological indicator. The OFC is located on the ventral surface of the prefrontal cortex and is associated with the integration of sensory information, memory, and decision-making [18]. More specifically, it plays an important role in regulating the cognitive and social behaviors related to rewards and punishments [19]. Additionally, the activity of the OFC has been shown to correlate with the pleasure or displeasure evoked by visual and olfactory stimuli, with activation of regions near the OFC when these emotions are evoked [20,21]. Furthermore, the left OFC is correlated with positive emotions, such as pleasure, while the right OFC is associated with negative emotions, such as displeasure, and a hemispheric difference in OFC responses has been reported [22]. Based on these previous studies, OFC responses could be considered physiological indicators for analyzing psychophysical experiments focusing on crossmodal perception between colors and fragrances.

NIRS is a non-invasive method that measures changes in the oxygenated hemoglobin (oxyHb) concentration using near-infrared light on biological tissues. This technique is particularly useful for measuring brain activity and is widely used in cognitive neuroscience and medical fields [23]. Near-infrared light is able to penetrate biological tissues such as the scalp, skull, and brain tissue. Additionally, hemoglobin in the blood has different absorption characteristics depending on its oxidation state. Based on these properties, light emitted from a source passes through the tissue, undergoes scattering and absorption within the body, and the reflected light is detected by a detector. The change in the intensity of the light is measured. This intensity change is based on the Beer–Lambert law. Another method for measuring brain activity is fMRI, which provides high spatial resolution by imaging not only structural information but also functional activity in the brain. However, fMRI requires large equipment and it is restrictive, as subject movement is not allowed. In contrast, NIRS, although having a lower spatial resolution than fMRI, has a low constraint, and brain function can be measured in more natural states or under light physical activities

without restrictions on the participant's posture. This makes NIRS a suitable method for the psychophysical experiments in this study. Therefore, in this study, we attempted to measure OFC responses using NIRS. During the experiment, NIRS (Hb133, ASTEM Co., Ltd., Kawasaki, Japan) was used to measure the oxyHb concentration changes at two locations above the eyebrows on the left and right sides (channel 1, near the left OFC, and channel 2, near the right OFC). The NIRS plate was positioned immediately above the eyebrows and symmetrically aligned. The total length of the plate was 14 cm, with the light source on the outside and the receiver placed on the inside. The measurement point corresponds to a 4 cm offset to the left or right from the center of the forehead. There are two channels, and they are not positioned in 3D. The sampling frequency of the measurement was 10 Hz, and during data analysis, a moving average was applied to the raw data. Changes in oxyHb concentrations were measured using the software supplied with Hb133. From this data, the difference in the peak values of the oxyHb concentrations before and during stimulus presentation (ΔoxyHb) was calculated, and this value was used as the response to the stimulus for analysis. The ΔoxyHb was calculated by the authors' own program based on those measurements.

2.2. Generation of Experimental Stimuli

First, we generated colored lights that served as visual stimuli for the participants. The colors used in the experiment were determined using a practical color coordinate system (PCCS), a two-dimensional system based on hue and tone that is suitable for considering color harmony. We referred to the vivid tone, which is the highest chroma tone in the PCCS, and selected 12 hues at equal intervals from the 24 vivid tones. Additionally, to ensure a sufficient examination of the colors that harmonize with fragrances, we prepared more colors by setting the chroma of the pure colors to two-thirds and one-third, respectively, based on white light (x = 0.33, y = 0.33). Consequently, we generated 36 colored lights by considering three chroma levels of the 12 hues. The chromaticity coordinates of the colored light are shown in Figure 1.

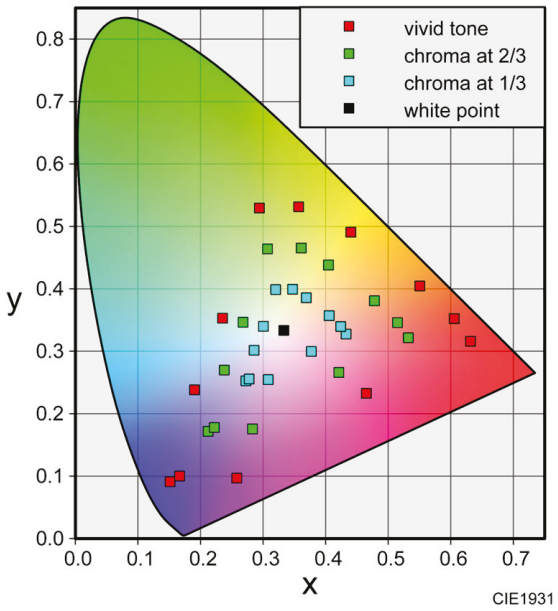


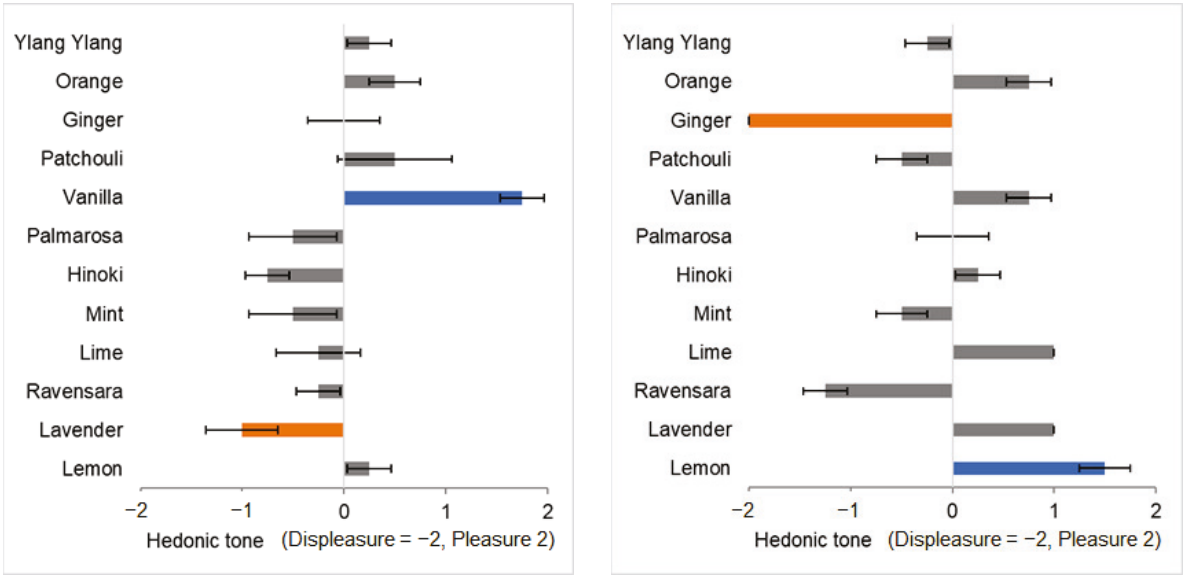
Figure 1. Chromaticity coordinates of the 36 generated colored lights.

Next, we generated fragrances that served as olfactory stimuli. We used essential oils that are considered suitable for eliciting physiological responses, owing to their ability to emit strong fragrances in small amounts. The five fragrances used by Yamashita et al. [16]

and Oba et al. [17]—vanilla, hinoki, mint, lime, and grapefruit—were relatively pleasant to the participants. In this study, we increased the variety of fragrances and analyzed the emotions evoked by unpleasant fragrances as well. We used 12 types of fragrances, including those used by Yamashita et al. [16] and Oba et al. [17], as well as unpleasant fragrances. The 12 fragrances used in the preliminary experiment were ylang-ylang, orange, ginger, patchouli, vanilla, palmarosa, hinoki, mint, lime, ravensara, lavender, and lemon. To minimize individual differences in the perception of pleasure or displeasure, we considered various categories such as citrus, floral, and resinous fragrances. Essential oils were absorbed on 1 cm^2 cotton pads and placed in 20 mL amber glass bottles. It was difficult to determine the color of the essential oils absorbed in these bottles. Olfactory stimuli were presented using the 'odor bottle method', in which participants brought the bottle close to their nose to smell the fragrance. The advantage of this method is that the bottles are lightweight and easy to hold, making it easy to present stimuli to participants and minimize movement, thereby reducing the noise in the NIRS measurements. Furthermore, we ensured that all fragrances had the same intensity by using a fragrance meter (OMX-SRM, Shinyei Technology Co., Ltd., Kobe, Japan) to adjust the fragrance intensity near the opening of the bottle to 100.

2.3. Selection Survey for Experimental Stimuli

We selected the appropriate experimental stimuli for each participant by conducting a preliminary survey with 36 colored lights and 12 fragrances. The survey was carried out with eight university students (age 23 $\pm$ 1 year). The participants were confirmed to have no olfactory abnormalities or familiarity with essential oils. Yamashita et al. [16] selected olfactory stimuli using a method in which participants selected the most pleasant or unpleasant fragrance among five fragrances. In this case, the pleasure or displeasure associated with each fragrance was relative, making it unclear whether the participants genuinely found the fragrance pleasant or unpleasant. Therefore, in this study, the participants scored each fragrance on a pleasant scale (hedonic tone) to achieve an evaluation that was as absolute as possible. Participants smelled each of the 12 fragrances and rated their hedonic tone on a subjective 5-point scale (-2 to 2). In addition, they described the type of fragrance they perceived. Scent selection data for all participants are shown in Figure S1 in the Supplemental Materials. Each participant completed the survey twice. To minimize the carryover effects from one fragrance to the next, a two-minute break was provided after each fragrance. As in the experiments by Yamashita et al. [16] and Oba et al. [17], participants were not informed of the names of the fragrances, which helped minimize the biases associated with the names and allowed them to evaluate pleasure or displeasure intuitively. Based on these evaluations, two fragrances were selected for each participant (one that they found the most pleasant and another that they found the most unpleasant). Figure 2 shows two examples of the fragrance evaluation results. The horizontal axis represents the pleasantness score, with '-2' indicating displeasure and '2' indicating pleasure. The error bars represent the standard error in the two trials. Next, we conducted a selection survey to determine harmonious and disharmonious color–fragrance combinations. The participants perceived each of the 36 colored lights while smelling each fragrance and rated the harmony between the colored light and the fragrance on an 11-point scale (-5 to 5). Based on these evaluations, we selected the 'most harmonious colored light (harmonious color)' and the 'most disharmonious colored light (disharmonious color)'. Consequently, two fragrances and four colored lights were selected as the experimental stimuli for each participant (Table 1).



(**a**) Participant 1 (**b**) Participant 2

Figure 2. Evaluation results of hedonic tone for fragrance. The pleasant fragrances selected are colored blue, and the unpleasant fragrances selected are colored orange. Also, (**a**) shows the results for participant 1, and (**b**) shows the results for participant 2. The error bars represent the standard error in the two trials.

Table 1. Type of selected stimulus.

Fragrances (2 Types)	Colored Lights (4 Types)
pleasant fragrance	harmonious color disharmonious color
unpleasant fragrance	harmonious color disharmonious color

There are several points to consider regarding harmony in this study. The first is the criteria that can be used to define what constitutes 'harmony'. The term 'harmony' has often been discussed in the context of intramodal relationships, such as the harmony among multiple sounds or among colors. While recent studies have increasingly focused on harmony across modalities, they have used different definitions of harmony. In this study, harmony was defined as the congruency between the fragrance and colored light. This definition was communicated to the participants before conducting the harmony-rating experiment. However, introspective reports revealed that when the names of the fragrances were not disclosed, participants sometimes failed to accurately imagine the nature of the fragrances, especially that of unpleasant fragrances. As a result, when the names of the fragrances were not provided, it became challenging to evaluate harmony based on the congruency between the fragrance and colored light. In such instances, the participants may have evaluated harmony more metaphorically, focusing on the smoothness and pleasantness of simultaneously perceiving fragrances and colored light. The second point concerns the characteristics of the participants. In this study, we confirmed that the participants were not familiar with essential oils and that some fragrances were difficult for them to identify. However, for individuals who regularly experience fragrances in their daily lives, it may be easier to form concrete imagery and perceive semantic congruency as harmonious, even without knowing the names of the fragrances. The third point pertains to how the mode of presentation of colored light affects the perception of harmony. The psychological states of the participants (such as their emotions) may vary depending on the environment

in which the colors are presented, which could influence their subjective evaluation of harmony. This may also apply when harmony is perceived metaphorically. Furthermore, lighting is perceived directly as light, whereas color patches are perceived as reflected light, indicating differences in the mode of perception that may impact the sense of harmony. Therefore, the way in which harmony is perceived may vary depending on the stimulus presentation method.

2.4. Experimental Environment

The participants sat on a chair with an observation box (approximately 45 cm wide and 55 cm high) placed in front of them (Figure 3) and illuminated from above using a spectrally adjustable lighting device (LEDCube, Thousand Lights Lighting (Changzhou) Limited, Changzhou, China). The LEDCube is a lighting environment simulation source capable of reproducing any color by adjusting the power output of monochromatic light. Their performance varied depending on the type of LED incorporated. In this experiment, 14 types of LEDs with different peak wavelengths were used, and the desired colored light was generated by entering the xy values into the software. A spectroradiometer (CL-500A, Konica Minolta, Inc., Tokyo, Japan) was used to verify that the appropriate colored light was emitted. The observation box and the participant were covered with dark curtains to minimize the influence of external stimuli. The illuminance within the observation box was set to 50 lx for all the colored lights, and the experiment was conducted in a dark room.

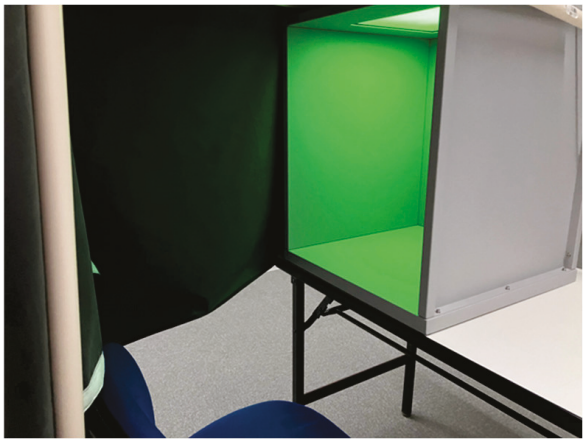


Figure 3. Lighting environment used in the experiment. This figure shows an example of green-colored light.

2.5. Experimental Procedure

NIRS was used to measure OFC activation in response to the presentation of colored lights and fragrances selected based on the selection survey. The experiment consisted of the four following conditions: (1) fragrance-only, (2) colored light-only, (3) harmonious crossmodal, and (4) disharmonious crossmodal presentation. The experimental results are summarized in Table 2. Eight university students participated in the selection survey. Each of the four types of experiments was conducted three times for all participants. Furthermore, since Yamashita et al. [16] suggested that the order of the four experimental types could influence ΔoxyHb, the order was randomized to prevent it from affecting the OFC responses. Approximately 30 to 60 s were allocated to each experiment prior to the stimulus presentation in order to stabilize the measurement equipment and brain activity. The stability of brain activity refers to the absence of changes in the oxyHb concentration near the OFC. After confirming stability, the stimuli were presented to the participants for 30 s. The presentation duration was determined based on Yamashita et al. [16], who suggested that a short duration of approximately 10 s did not show a stimulus response,

whereas a duration that was too long could lead to reduced participant concentration, thus affecting brain activity. After the presentation, the participants rested for approximately 60 s, during which the response progression was fully monitored. Breaks were taken between presentations, and the well-being of the participants was monitored throughout the experiment.

Table 2. Breakdown of the four types of experiments.

Experiment (1): Fragrance-Only	Experiment (2): Colored Light-Only	Experiment (3): Harmonious Crossmodal	Experiment (4): Disharmonious Crossmodal
pleasant fragrance	harmonious color disharmonious color	pleasant–armonious	pleasant–disharmonious
unpleasant fragrance	harmonious color disharmonious color	unpleasant–harmonious	unpleasant–disharmonious

The experimental procedures for each condition are summarized below.

(1) Fragrance-only
A sniffing test was conducted using odor bottles. After 30 s of sniffing, the participants were asked to move the bottle away from their nose and return it to the position at their chest. Then, the participants were observed for at least 30 s until their responses stabilized.

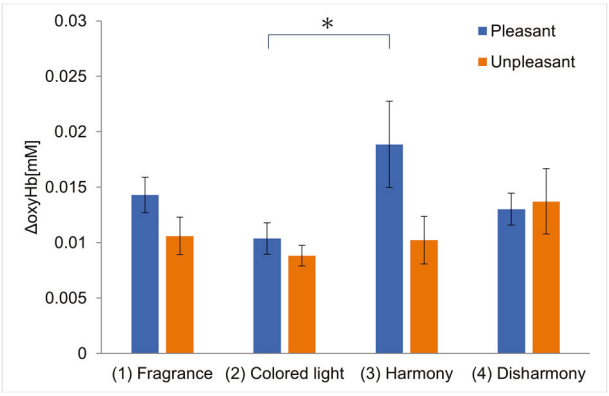
(2) Colored light-only
There was no task for the participants to perform. As soon as the NIRS equipment was stabilized, the experimenter changed the lighting color from white to the desired color, and the participants perceived that color for 30 s. After 30 s, the color was reverted to white. Then, the participants were observed for at least 30 s until their responses stabilized.

(3) Harmonious and (4) disharmonious crossmodal
A sniffing test using odor bottles was used for the presentation of aroma stimuli. The participant held a bottle containing the fragrance near his/her chest before the stimulus presentation. As soon as the NIRS device was stabilized, the participant brought the bottle close to his/her nose and sniffed the fragrance upon the experimenter's cue. After 30 s, the experimenter cued the participants to move the bottle away from their nose and change the color back to white. Then, the participants were observed for at least 30 s until their responses stabilized.

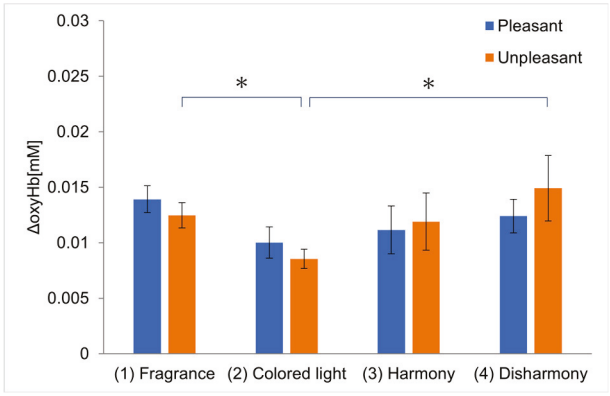
2.6. Results and Discussion

2.6.1. Difference in Peak Values of oxyHb Concentration in the Left and Right Orbitofrontal Cortex

We measured OFC activation under four types of stimulus presentations using NIRS. Figure 4 shows the oxyHb levels in the left and right OFC for each experiment. The vertical axis represents the average ΔoxyHb across all participants, and the horizontal axis compares the four experimental conditions. For the colored light-only presentation, the ΔoxyHb values represent the average of the harmonious and disharmonious colors. The error bars indicate the standard error of the mean ΔoxyHb across participants. Additionally, a Smirnov–Grubbs test at a 5% significance level was conducted on all measurement data to remove the outliers, which comprised two data points (approximately 2% of the total data). Figures 4a and 4b show the results for channel 1 and channel 2, representing the left and right OFC, respectively.



(**a**) ΔoxyHb in channel 1 (left OFC)
p_p = 0.02 (η_p^2 = 0.41), p_u = 0.27 (η_p^2 = 0.19)



(**b**) ΔoxyHb in channel 2 (right OFC)
p_p = 0.24 (η_p^2 = 0.14), p_u = 0.08 (η_p^2 = 0.30)

Figure 4. Comparison of ΔoxyHb in the four types of experiments. The asterisks indicate a significance level of 5% in the *t*-test, and the error bars represent the standard error of the mean ΔoxyHb among participants. The same applies to Figures 5–7. The error bars indicate the standard error of the mean ΔoxyHb across participants.

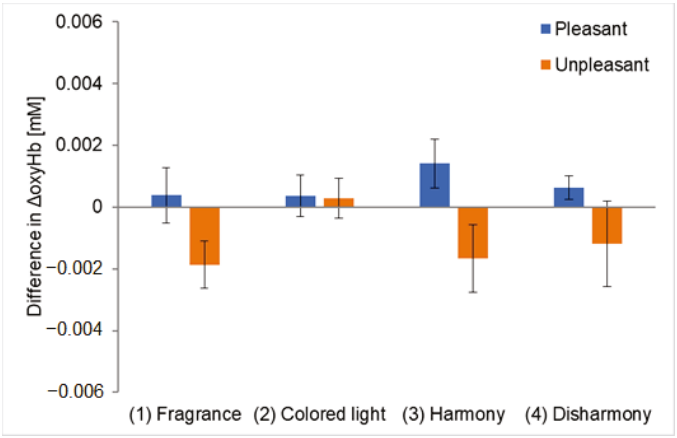
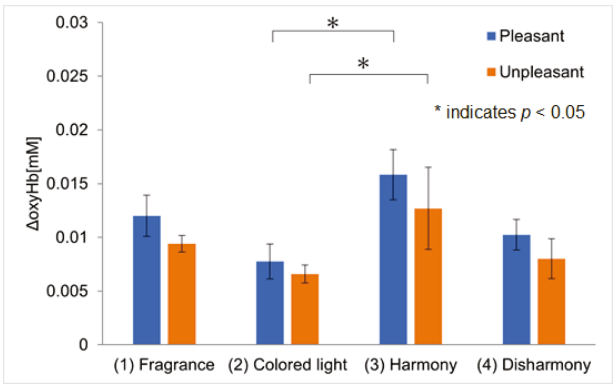
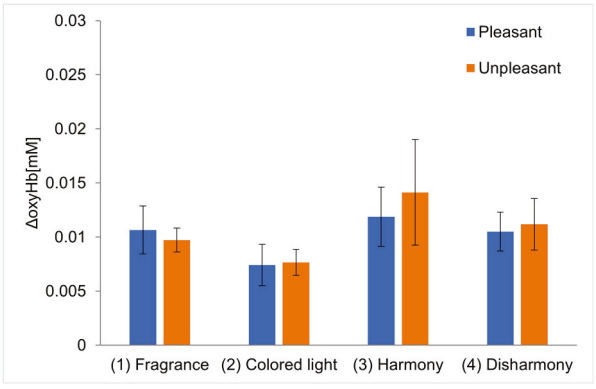


Figure 5. The difference in ΔoxyHb between the left and right OFC. A positive value indicates greater activation in the left OFC, whereas a negative value indicates activation in the right OFC. The error bars represent the standard error of the mean difference in ΔoxyHb across participants.



(**a**) ΔoxyHb in channel 1 (left OFC)
p_p = 0.00 (η_p^2 = 0.49), p_u = 0.18 (η_p^2 = 0.20)



(**b**) ΔoxyHb in channel 2 (right OFC)
p_p = 0.23 (η_p^2 = 0.18), p_u = 0.22 (η_p^2 = 0.18)

Figure 6. Comparison of ΔoxyHb between four experimental conditions (in Experiment II). The error bars indicate the standard error of the mean ΔoxyHb of the participants.

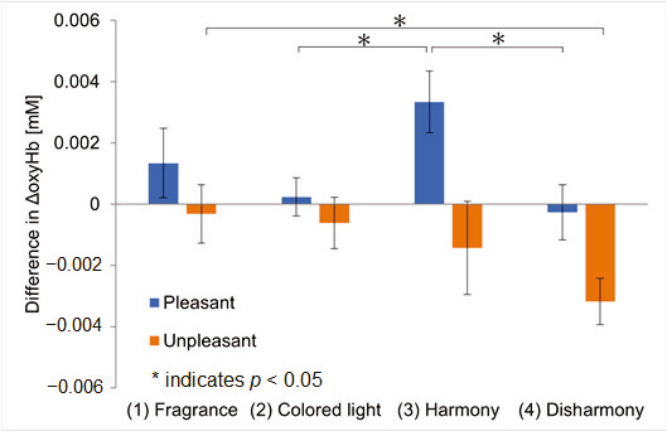


Figure 7. The difference in ΔoxyHb between the left and right OFC (in Experiment II). The error bars represent the standard error of the mean ΔoxyHb difference among the participants.

For the pleasant fragrance in channel 1, ΔoxyHb was highest during the crossmodal presentation with the harmonious color (Figure 4a). This suggests that the participants experienced greater pleasure when perceiving the fragrance with its harmonious color than when perceiving the fragrance alone. A statistically significant difference was found in OFC activation during the colored light-only presentation and the harmonious crossmodal presentation (* indicates $p < 0.05$). Meanwhile, for the unpleasant fragrance, ΔoxyHb was higher during the crossmodal presentation with the disharmonious color, although the effect was not as pronounced as with the pleasant fragrance. No statistically significant differences were observed among combinations that used unpleasant fragrances. Furthermore, when comparing ΔoxyHb for pleasant and unpleasant fragrances in the four experimental conditions, a substantial difference was observed in the harmonious crossmodal presentation, supporting the inference that the left OFC is associated with pleasant emotions.

For the pleasant fragrance in channel 2, ΔoxyHb increased slightly in the fragrance-only presentation, although no significant differences were observed in the comparisons made with other conditions (Figure 4b). No statistically significant differences were found in channel 2 for the pleasant fragrances for any combination. However, for the unpleasant fragrance, ΔoxyHb was the highest during the crossmodal presentation with the disharmonious color. This suggests that perceiving an unpleasant fragrance with a disharmonious color intensified the unpleasant experience compared with perceiving the fragrance alone. Statistically significant differences were observed between the colored light-only and disharmonious crossmodal presentations and between the fragrance-only and disharmonious crossmodal presentations. Furthermore, the difference in the ΔoxyHb for pleasant and unpleasant fragrances was highest in the disharmonious crossmodal presentation, suggesting that the right OFC is associated with unpleasant emotions. These results differ from those of Yamashita et al. [16], who found that ΔoxyHb was highest when the fragrance was presented alone, suggesting that the order in which the experimental stimuli were presented may influence the findings. Furthermore, the ΔoxyHb results in the colored light-only presentation suggest that perceiving light alone does not activate the OFC as strongly as the fragrance-only or crossmodal presentation, indicating a strong relationship between the OFC and emotions evoked by fragrance.

We have calculated a repeated-measures ANOVA and added values to the captions of Figure 4a,b. Because this study claims that the effect is obtained for a particular combination of two factors, the ANOVA analysis is only a reference, but the numbers generally support the claim in terms of p-values and partial η squared.

2.6.2. Differences in Difference in Peak Values of oxyHb Concentration Between the Left and Right Orbitofrontal Cortex

Conventional studies have suggested that the left and right OFC may play distinct roles in processing multisensory integration, particularly in evaluating congruent and incongruent sensory inputs [24,25]. Stronger activation of the left OFC under harmonious crossmodal conditions could be related to its role in reward processing and integration of pleasant sensory experiences. The left OFC has been associated with positive affective processing, which could explain its enhanced activity in response to harmonious stimuli.

Conversely, the right OFC's stronger activation under disharmonious conditions may reflect its involvement in conflict detection, aversive processing, or the evaluation of unexpected stimuli. The right OFC has been linked to the detection of violations of expectations and the processing of conflicting sensory inputs, which may explain its increased activation under disharmonious conditions.

Figure 5 shows the differences in ΔoxyHb between the left and right OFC. The vertical axis represents the differences in ΔoxyHb between channel 1 and channel 2, calculated using Equation (1):

$$Difference\ in\ \Delta oxyHb = (\Delta oxyHb\ in\ channel1) - (\Delta oxyHb\ in\ channel2) \quad (1)$$

A positive value indicates greater activation in channel 1, whereas a negative value indicates greater activation in channel 2. The error bars represent the standard error of the mean difference in ΔoxyHb across participants. As in the ΔoxyHb analysis in Section 2.6.1, outliers were excluded from this analysis as well. For pleasant fragrances, the crossmodal presentation with harmonious colors resulted in the largest channel difference, indicating greater activation of the left OFC. This difference suggests that pleasant emotions were strongest during the harmonious crossmodal presentation, although no statistically significant differences were observed. For unpleasant fragrances, the fragrance-only and crossmodal presentations showed larger response differences, indicating greater activation of the right OFC during these presentations. However, the standard error was higher for the disharmonious crossmodal presentation.

The differences in the patterns observed for pleasant and unpleasant fragrances may be related to participants' impressions of the fragrances in the selection survey (Section 2.3). As mentioned in Section 2.3, the participants reported being able to imagine an object associated with pleasant fragrances, while impressions of unpleasant fragrances were often vague and unclear. Specifically, although the names of the fragrances were not provided, pleasant fragrances such as orange and lemon often evoked clear mental images, whereas unpleasant fragrances such as ginger and ravensara did not. This suggests that the lack of clarity in participants' perceptions of unpleasant fragrances may have led to a less accurate matching of harmonious and disharmonious colors with unpleasant fragrances than with pleasant fragrances. This experiment is based on the sensory congruence between fragrance and lighting color, independent of the bias caused by the names of the fragrances, in the analysis of pleasant and unpleasant feelings. However, it is important to note that the degree of imagery evoked by fragrances differs between pleasant and unpleasant odors. Therefore, it is possible that this influenced the OFC's activation. Additionally, the differences between channels observed during the fragrance-only presentations suggest that displeasure may have intensified because of the participants' inability to imagine the unpleasant fragrance. Furthermore, the increased standard error during the crossmodal presentations could be due to variations in how the participants perceived harmony with unpleasant fragrances. Most participants reported feeling 'more pleasant' when a harmonious color was presented with a pleasant fragrance and 'more unpleasant' when a disharmonious color was presented. For unpleasant fragrances, they felt 'more pleasant' with a harmonious color

and 'more unpleasant' with a disharmonious color. However, one participant out of eight reported feeling 'more unpleasant' with a harmonious color for the unpleasant fragrance, which may have contributed to the increased error. In this experiment, essential oils were used as olfactory stimuli, which may explain the observed tendency toward judging the fragrances as unpleasant. However, for fragrances that were more unpleasant and easily imaginable by participants, even when presented with a harmonious color, it is possible that they perceived them as unpleasant.

3. Experiment II: Analysis of Functional Differences Between the Left and Right Orbitofrontal Cortex

In Experiment I, we conducted the study based on previous findings [20–22] suggesting that the left OFC is associated with pleasant emotions, while the right OFC is associated with unpleasant emotions. Our results indicated that in the crossmodal presentation of a pleasant fragrance and its harmonious color, activation of the left OFC enhanced pleasant emotions. Conversely, in the crossmodal presentation of an unpleasant fragrance and its disharmonious color, activation of the right OFC may have enhanced unpleasant emotions. A recent crossmodal study [23] has presented differing views on the functional differences between the left and right OFC in relation to pleasure and displeasure. Specifically, the left OFC is activated when the interaction between colored light and fragrance becomes more harmonious, promoting a crossmodal interaction and becoming dominant in crossmodal perception. In contrast, the right OFC shows a smaller effect from the interaction and is activated when the stimuli are perceived independently, in a more multimodal perception, highlighting this functional difference. Furthermore, studies have shown that even without directly smelling a fragrance, imagining the fragrance or recalling olfactory experiences can activate sensory regions related to olfactory perception, with stronger activation observed in the left OFC [26,27]. Additionally, Broman et al. [28] showed that the right hemisphere is involved in low-level olfactory processing and encoding based on perception, whereas the left hemisphere is associated with higher-order olfactory recognition and semantic interpretation, such as memory and the labeling of fragrances. These findings suggest that the functional differences between the left and right OFC may not be solely determined by 'pleasure–displeasure'. Instead, another factor may be the extent to which crossmodal perception is enhanced by harmony (higher-order modality-integrated perception), as opposed to stimuli being merely perceived independently due to disharmony (lower-order modality-integrated perception or multimodal perception). In Experiment I, participants were not informed of the names of the fragrances during the stimulus selection phase. However, it has been reported that pleasant fragrances tend to be easier to associate with their corresponding objects, whereas unpleasant fragrances are more difficult to imagine. Given the findings [23,26–28], it is possible that this difference in imagery influenced the crossmodal perception of colored light and fragrance. Therefore, in Section 3, we conduct a new NIRS experiment using fragrances that are familiar and easy to imagine for all participants, ensuring that both pleasant and unpleasant fragrances are equally accessible in terms of mental imagery. By analyzing the results, we aim to analyze whether the functional differences between the left and right OFC are influenced not only by the 'pleasure–displeasure axis' but also by the 'crossmodal perception–multimodal perception axis'.

3.1. Selection of Experimental Stimuli and Experimental Procedure

We selected the same stimuli as in Experiment I. The same participants (age 23 ± 1 year) participated in this experiment. We selected familiar fragrances from the 12 essential oils used in Experiment I. Eight fragrances were selected: orange, ginger, vanilla, hinoki, mint, lime, lavender, and lemon. The 'most pleasant' and 'most unpleasant'

fragrances were selected from this list following the same procedure as in Section 2.3. Scent selection data for all participants are shown in Figure S2 in the Supplemental Materials. This survey revealed that knowledge of the name of the fragrance affected the subjective hedonic tone judgment of many participants. This finding indicates that the imagery evoked by the names of fragrances significantly influences the pleasant–unpleasant evaluation of the fragrance. Harmonious and disharmonious colors were identified for the selected fragrances. As in Experiment I, the participants were informed of the definition of harmony prior to performing the evaluation experiment using the same experimental conditions as described in Section 2.5. However, because they were aware of the names of the fragrances, we inferred that the participants evaluated harmony based on the semantic congruency between the fragrance and colored light.

3.2. Results and Discussion

3.2.1. Difference in Peak Values of oxyHb Concentration in the Left and Right Orbitofrontal Cortex

NIRS was used to measure OFC activation in response to the four types of presentations. The oxyHb values for each experiment are shown in Figure 6; the vertical axis represents the average ΔoxyHb across all participants, while the error bars indicate the standard error of the mean ΔoxyHb of the participants. The ΔoxyHb for the colored light-only presentation is the average value obtained from responses to harmonious and disharmonious colors. Five outliers were removed using the Smirnov–Grubbs test at a 5% significance level, accounting for approximately 5% of the total data. Figures 6a and 6b show the analysis results for channel 1 and channel 2, corresponding to the left and right OFC, respectively. The ΔoxyHb measured on channel 1 was the highest during the crossmodal presentations of pleasant fragrances with harmonious colors, similar to the results of Experiment I (Figure 6a). However, ΔoxyHb was also high during the crossmodal presentations of unpleasant fragrances with harmonious colors, a result different from that seen in Experiment I. This suggests that the participants might have imagined harmony more effectively for unpleasant fragrances in Experiment II, leading to dominant crossmodal perception and activation in the left OFC. Statistically significant differences were observed between the colored light-only and harmonious crossmodal presentations for both pleasant and unpleasant fragrances. For channel 2, while ΔoxyHb was slightly higher during the harmonious crossmodal presentation with pleasant fragrances, the difference was not significant (Figure 6b). For unpleasant fragrances, ΔoxyHb peaked during a harmonious crossmodal presentation; however, in contrast to Figure 6a, the error was large, and no statistically significant differences were observed between conditions. We have calculated a repeated-measures ANOVA and added values to the captions of Figure 6a,b.

3.2.2. Differences in Difference in Peak Values of oxyHb Concentration Between the Left and Right Orbitofrontal Cortex

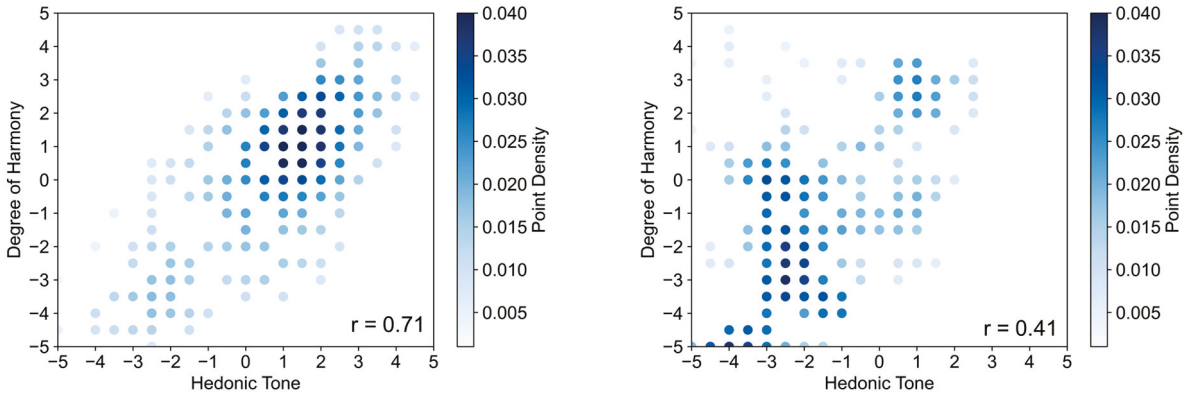
Figure 7 shows the results of the analysis of the ΔoxyHb differences between the left and right OFC, the vertical axis representing the difference in ΔoxyHb between channel 1 and channel 2, calculated using Equation (1). The error bars represent the standard error of the mean ΔoxyHb difference among the participants. The outliers described in Section 3.2.1 were excluded from this analysis as well. For pleasant fragrances, the largest channel difference was observed during the crossmodal presentations with harmonious colors, confirming greater activation in the left OFC. This suggests that the greatest sense of pleasure was evoked during the crossmodal presentation with harmonious colors, yielding results similar to those of Experiment I. However, in contrast to Experiment I, significant differences were found between the 'colored light-only and harmonious crossmodal' and 'harmonious and disharmonious crossmodal' presentations, suggesting that knowing the

names of the fragrances may enhance pleasure or create dominant crossmodal effects. For unpleasant fragrances, the difference in responses increased during the disharmonious crossmodal presentations, with greater activation in the right OFC. This result differs from that of Experiment I, indicating that the participants might have perceived unpleasant fragrances and disharmonious colored light as individual stimuli, leading to the activation of the right OFC. Therefore, Experiment II suggests that the left and right OFC were influenced not only by a 'pleasure–displeasure axis' but also by a 'crossmodal perception–multimodal perception axis'.

4. Additional Experiment: Correlation Between Hedonic Tone and Degree of Harmony in Subjective Evaluation

The results of Experiment I suggest that when a pleasant fragrance and its harmonious color are crossmodally perceived, the left OFC is activated, potentially enhancing pleasant emotions. In contrast, when an unpleasant fragrance and its disharmonious color are crossmodally perceived, the right OFC is activated, potentially enhancing unpleasant emotions. In our experiments in the previous sections, NIRS was used to objectively analyze the recall of pleasant and unpleasant emotions while considering the functional differences of the left and right OFC, as reported in the literature [20–22]. However, it was not possible to verify whether the participants judged pleasure or displeasure subjectively. Therefore, as an additional experiment, the relationship between the subjective hedonic tone and the degree of harmony in the crossmodal perception of colored light and fragrance was investigated.

First, an experiment was conducted where the subjective hedonic tone of the crossmodal perception of two types of pleasant and unpleasant fragrances selected in 3.2 and 36 lighting colors generated in 3.1 was evaluated. There were six of the eight participants who participated in Experiment I (age 23 $\pm$ 1 year). The participants simultaneously perceived the colored light and fragrance and rated the hedonic tone of the experience on an 11-point scale. The experimental environment was the same as in Section 2.4. Using the evaluation results of hedonic tone and the degree of harmony evaluation results from Section 2.3, a scatter plot showing their correlation is shown in Figure 8a shows the results for the pleasant fragrance, and (b) shows the results for the unpleasant fragrance. The vertical axis represents the degree of harmony, and the horizontal axis represents the hedonic tone. Additionally, the closer the score is to '5', the more harmonious or pleasant it is, while the closer it is to '−5', the more disharmonious or unpleasant it is. The correlation coefficients for each are noted in the lower right of the scatter plots. The density of points due to overlapping is shown using a heatmap based on Kernel Density Estimation.



(**a**) pleasant fragrance (**b**) unpleasant fragrance

Figure 8. Correlation between hedonic tone and degree of harmony.

In the case of the pleasant fragrance shown in Figure 8a, a high positive correlation of 0.71 was confirmed between the hedonic tone and harmony in the crossmodal perception of colored light and fragrance. This indicates that for the pleasant fragrance, colored lights with higher harmony were perceived as more pleasant, while those with lower harmony suppressed or induced displeasure. This result supports the findings of Experiment I, where the left OFC was activated when a pleasant fragrance and its harmonious color were crossmodally perceived, enhancing pleasant emotions. In contrast, for the unpleasant fragrance shown in Figure 8b, the correlation coefficient was 0.41, which is lower than for the pleasant fragrance but still indicates a positive correlation. This suggests that for the unpleasant fragrance, colored lights with lower harmony were perceived as more unpleasant, while those with higher harmony suppressed or induced pleasure. This result also supports the findings of Experiment I, where the right OFC was activated when an unpleasant fragrance and its disharmonious color were crossmodally perceived, enhancing unpleasant emotions. The slightly lower correlation for the unpleasant fragrance can be attributed to the difference in how one participant perceived harmony, as mentioned in Section 2.6.2. This participant perceived the combination of the unpleasant fragrance and its harmonious color as more unpleasant, which may have influenced the negative correlation. When excluding this participant, the correlation coefficient was 0.55.

In Experiment II, it was suggested that the functional differences between the left and right OFC are also influenced by the 'crossmodal perception–multimodal perception axis'. Focusing on the degree of harmony between the colored light and fragrance, it was found that when the degree of harmony is high, semantic crossmodal perception becomes dominant, resulting in higher-order modality-integrated perception. In contrast, when the degree of harmony is low, the stimuli are not semantically linked and are processed merely as independent stimuli, resulting in lower-order modality-integrated perception. Based on this, it can be inferred that for the pleasant fragrance, higher harmony in the colored light enhances pleasant emotions, indicating that crossmodal perception leads to pleasure. Conversely, for the unpleasant fragrance, lower harmony in the colored light enhances unpleasant emotions, suggesting that multimodal perception leads to displeasure. Therefore, it can be concluded that the 'pleasant–unpleasant axis' and the 'crossmodal perception–multimodal perception axis' are interrelated in the functional differences between the left and right OFC.

The difference between crossmodal and multimodal perception can be viewed neuroscientifically as a difference in the following actions [29].

Crossmodal perception involves interactions between different sensory modalities, wherein information from one sense influences the perception of another. This process may engage multiple brain regions, such as:

- The orbitofrontal cortex and insula, which integrate sensory inputs to form coherent perceptual experiences.
- The superior temporal sulcus, which plays a key role in integrating audiovisual stimuli.
- The intraparietal sulcus, which coordinates attentional resources across the senses.

The piriform cortex and visual association area information are linked, and the sense of smell is enhanced by vision. Crossmodal processing might be characterized by dynamic interactions between these regions and often requires more extensive neural integration than unimodal processing.

By contrast, multimodal perception refers to the parallel processing of multiple sensory inputs without necessarily integrating them into a unified experience. This process often relies more on independent sensory-specific brain regions. Primary sensory cortices (e.g., the visual, entorhinal, and piriform cortices) process different sensory inputs separately.

Higher-order association areas may still interact, but the level of integration is lower than that of crossmodal perception.

5. Conclusions

This study focused on the lateralization of brain function in the left and right OFC by analyzing the pleasure–displeasure evoked by harmony between colored light and fragrance. Two experiments were conducted, each following the processes of stimulus generation, selection, and NIRS measurement during stimulus presentation. In Experiment I, 36 colored lights and 12 fragrances were used as stimuli. A selection survey was used to have participants rate the hedonic tones of 12 fragrances without knowing their names—of which the most pleasant and unpleasant fragrances were selected. The participants then rated the degree of harmony between the selected fragrances and colored lights, leading to the selection of 'harmonious' and 'disharmonious' color–fragrance combinations. Each participant was presented with two different fragrances and four colored lights as the experimental stimuli. Based on these selected stimuli, NIRS measurements of OFC activity were performed under the four following presentation conditions: fragrance-only, colored light-only, harmonious crossmodal, and disharmonious crossmodal presentations. The NIRS device measured changes in the oxyHb concentrations in the left and right OFC. Moving averages were applied to the measured data, and the difference in oxyHb concentration (ΔoxyHb) before and after stimulus presentation was calculated as a response to each stimulus. We found that the ΔoxyHb associated with channel 1 was the largest during the crossmodal presentation with harmonious colors. This suggests that the participants felt more pleasant when they perceived the harmonious color simultaneously with the pleasant fragrance rather than when only smelling the pleasant fragrance. The ΔoxyHb measured on channel 2 was the largest during the crossmodal presentation with disharmonious colors. This suggests that participants felt more displeasure when they perceived the disharmonious color simultaneously with the unpleasant fragrance rather than when they only smelled the unpleasant fragrance. In Experiment II, eight familiar fragrances were selected from the original list of twelve fragrances, and the same procedures as in Experiment I were followed to select the 'most pleasant' and 'most unpleasant' fragrances, as well as the 'harmonious' and 'disharmonious' colors associated with them. The NIRS measurements in Experiment II showed that ΔoxyHb was highest in channel 1 for the pleasant and unpleasant fragrances in the harmonious crossmodal presentation, suggesting that the left OFC is more active when crossmodal effects are dominant. An additional experiment showed a positive correlation between hedonic tones and the degree of harmony, indicating that harmonious colors evoke pleasure, whereas disharmonious colors evoke displeasure. This suggests that the lateralization of brain function in the left and right OFC arises from the interaction between the 'pleasure–displeasure axis' and 'crossmodal perception–multimodal perception axis'.

Although our work is rooted in experimental psychophysics, its significance extends to the field of sensing technology, particularly to the application and optimization of NIRS for studying multisensory interactions. To better highlight this aspect, we have refined our discussion to emphasize the following points:

1. Innovation in Sensing Methodology

Our study employed NIRS in a novel way to examine the psychophysical correlates of crossmodal interactions between light and olfactory stimuli. Unlike conventional applications of NIRS, we specifically investigated its use for detecting subtle hemodynamic responses associated with sensory integration.

2. Contribution to Sensor-Based Research

Although we did not design a new sensor, our approach introduces an innovative application of existing sensing technology to a relatively unexplored research domain. By demonstrating how NIRS can be effectively utilized in multisensory studies, our study lays the groundwork for further developments in sensor-based neuroscience and biomedical engineering.

Future studies should explore different methods for analyzing OFC responses. In this study, moving averages were applied to the measurement data, and oxyHb was used as the response to each stimulus. Further analyses that focus on deoxygenated hemoglobin using GLM-based analyses or a combination of these methods may allow for a more objective analysis of the physiological responses. Furthermore, the fragrances used in this study were essential oils, which may have limited the range of displeasure tested in this analysis. Using more intense and unpleasant fragrances could potentially reveal a more distinct lateralization of brain function in the OFC. Moreover, as mentioned in Section 2.3, differences in the color presentation method may influence the harmony perceived by participants; this requires experimental demonstration. While the sample size of eight participants is relatively small, it is required for larger follow-up studies to validate our results. Although gender differences were not a primary focus of this study, we recognize their potential influence on sensory perception and neural responses. Also, future research should incorporate a more diverse sample to examine potential gender-related variations. A recent study has proposed a method for understanding odor descriptions in multimodal contexts, where odor information is communicated indirectly through text and images [30]. An analysis of the relevance of those multimodal studies is also an issue.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/s25072230/s1, A Results of Fragrance Selection Survey for all participants.

Author Contributions: Conceptualization, T.O. and T.H.; methodology, T.O. and T.H.; software, T.O.; validation, T.O., M.T. and T.H.; formal analysis, T.O.; investigation, T.O., M.T. and T.H.; resources, T.H.; data curation, T.H.; writing—original draft preparation, T.O.; writing—review and editing, M.T. and T.H.; visualization, T.O.; supervision, T.H.; project administration, T.H.; funding acquisition, T.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by JSPS KAKENHI, grant number 20H05957 and 24K02954.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study. Ethical review and approval were waived for this study on the grounds that the stress caused to the participants by the experiment was of the magnitude of everyday life.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study, in the collection, analyses, or interpretation of data, in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

OFC	Orbitofrontal Cortex
NIRS	Near-Infrared Spectroscopy
fMRI	functional Magnetic Resonance Imaging
PCCS	Practical Color Coordinate System
oxyHb	Oxygenated Hemoglobin

References

1. Spence, C. Olfactory-colour crossmodal correspondences in art, science, and design. *Cogn. Res. Princ. Implic.* **2020**, *5*, 52. [CrossRef] [PubMed]
2. Shams, L.; Kim, R. Crossmodal influences on visual perception. *Phys. Life Rev.* **2010**, *7*, 269–284. [CrossRef] [PubMed]
3. Zellner, D.A.; Kautz, M.A. Color affects perceived odor intensity. *J. Exp. Psychol. Hum. Percept. Perform.* **1990**, *16*, 391–397. [CrossRef]
4. Gilbert, A.N.; Martin, R.; Kemp, S.E. Cross-Modal Correspondence between Vision and Olfaction: The Color of Smells. *Am. J. Psychol.* **1996**, *109*, 335–351. [CrossRef]
5. Miura, K.; Saito, M. Harmonious color model with fragrances. *Color Res. Appl.* **2012**, *37*, 219–232. [CrossRef]
6. Demattè, M.L.; Sanabria, D.; Spence, C. Cross-Modal Associations Between Odors and Colors. *Chem. Senses* **2006**, *31*, 531–538. [CrossRef]
7. Demattè, M.L.; Sanabria, D.; Spence, C. Olfactory Discrimination: When Vision Matters? *Chem. Senses* **2009**, *34*, 103–109. [CrossRef]
8. Österbauer, R.A.; Matthews, P.M.; Jenkinson, M.; Beckmann, C.F.; Hansen, P.C.; Calvert, G.A. Color of Scents: Chromatic Stimuli Modulate Odor Responses in the Human Brain. *J. Neurophysiol.* **2005**, *93*, 3434–3441. [CrossRef] [PubMed]
9. Pinti, P.; Aichelburg, C.; Gilbert, S.; Hamilton, A.; Hirsch, J.; Burgess, P.; Tachtsidis, I. A Review on the Use of Wearable Functional Near-Infrared Spectroscopy in Naturalistic Environments. *Jpn. Psychol. Res.* **2018**, *60*, 347–373. [CrossRef]
10. Quaresima, V.; Ferrari, M. A Mini-Review on Functional Near-Infrared Spectroscopy (fNIRS): Where Do We Stand, and Where Should We Go? *Photonics* **2019**, *6*, 87. [CrossRef]
11. Han, P.; Zang, Y.; Akshita, J.; Hummel, T. Magnetic Resonance Imaging of Human Olfactory Dysfunction. *Brain Topogr.* **2019**, *32*, 987–997. [CrossRef] [PubMed]
12. Zhou, G.; Lane, G.; Cooper, S.L.; Kahnt, T.; Zelano, C. Characterizing functional pathways of the human olfactory system. *eLife* **2019**, *8*, e47177. [CrossRef]
13. Plailly, J.; Delon-Martin, C.; Royet, J.-P. Experience Induces Functional Reorganization in Brain Regions Involved in Odor Imagery in Perfumers. *Hum. Brain Mapp.* **2012**, *33*, 224–234. [CrossRef] [PubMed]
14. Zellner, D.A.; Whitten, L.A. The Effect of Color Intensity and Appropriateness on Color-Induced Odor Enhancement. *Am. J. Psychol.* **1999**, *112*, 585–604. [CrossRef]
15. Porada, D.K.; Regenbogen, C.; Seubert, J.; Freiherr, J.; Lundström, J.N. Multisensory Enhancement of Odor Object Processing in Primary Olfactory Cortex. *Neuroscience* **2019**, *418*, 254–265. [CrossRef]
16. Yamashita, G.; Tanaka, M.; Horiuchi, T. Analysis of harmony between color and fragrance in lighting environments by the reaction of the orbitofrontal area. *i-Perception* **2022**, *13*, 1–13. [CrossRef]
17. Oba, T.; Tanaka, M.; Horiuchi, T. Analysis of harmony between coloured light and fragrance by the reaction of the left-right orbitofrontal area. In Proceedings of the Congress of the International Colour Association, Chiang Rai, Thailand, 27 November–1 December 2023; pp. 1045–1050.
18. Kringelbach, M.L. The human orbitofrontal cortex: Linking reward to hedonic experience. *Nat. Rev. Neurosci.* **2005**, *6*, 691–702. [CrossRef]
19. Bechara, A.; Damasio, A.R.; Damasio, H.; Anderson, S.W. Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition* **1994**, *50*, 7–15. [CrossRef] [PubMed]
20. Noriuchi, M.; Kikuchi, Y.; Senoo, A. The Functional Neuroanatomy of Maternal Love: Mother's Response to Infant's Attachment Behaviors. *Biol. Psychiatry* **2008**, *63*, 415–423. [CrossRef]
21. Royet, J.-P.; Plailly, J.; Delon-Martin, C.; Kareken, D.A.; Segebarth, C. fMRI of emotional responses to odors:: Influence of hedonic valence and judgment, handedness, and gender. *NeuroImage* **2003**, *20*, 713–728. [CrossRef]
22. Gottfried, J.A.; Dolan, R.J. The Nose Smells What the Eye Sees: Crossmodal Visual Facilitation of Human Olfactory Perception. *Neuron* **2003**, *39*, 375–386. [CrossRef]
23. Boot, E.; Levy, A.; Gaeta, G.; Gunasekara, N.; Parkkinen, E.; Kontaris, E.; Jacquot, M.; Tachtsidis, I. fNIRS a novel neuroimaging tool to investigate olfaction, olfactory imagery, and crossmodal interactions: A systematic review. *Front. Neurosci.* **2024**, *18*, 1266664. [CrossRef]
24. Rolls, E.T. The orbitofrontal cortex and reward. *Cereb Cortex.* **2000**, *10*, 284–294. [CrossRef] [PubMed]
25. Rolls, E.T.; Cheng, W.; Feng, J. The orbitofrontal cortex: Reward, emotion and depression. *Brain Commun.* **2020**, *2*, fcaa196. [CrossRef] [PubMed]
26. Djordjevic, J.; Zatorre, R.J.; Petrides, M.; Boyle, J.A.; Jones-Gotman, M. Functional neuroimaging of odor imagery. *NeuroImage* **2005**, *24*, 791–801. [CrossRef]
27. Bensafi, M.; Sobel, N.; Khan, R.M. Hedonic-Specific Activity in Piriform Cortex During Odor Imagery Mimics That During Odor Perception. *J. Neurophysiol.* **2007**, *98*, 3254–3262. [CrossRef] [PubMed]

28. Broman, D.A.; Olsson, M.J.; Nordin, S. Lateralization of Olfactory Cognitive Functions: Effects of Rhinal Side of Stimulation. *Chem. Senses* **2001**, *26*, 1187–1192. [CrossRef]
29. Teichert, M.; Bolz, J. How Senses Work Together: Cross-Modal Interactions between Primary Sensory Cortices. *Neural Plast.* **2018**, *2018*, 5380921. [CrossRef]
30. Sergio, E.-R.; Iván, M.-F.; Manuel, G.-M.; Fernando, F.-M. Synthesizing Olfactory Understanding: Multimodal Language Models for Image-Text Smell Matching. 2024. Available online: https://ssrn.com/abstract=4912100 (accessed on 28 March 2025).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

sensors

MDPI

Article

Parameter-Efficient Adaptation of Large Vision—Language Models for Video Memorability Prediction

Iván Martín-Fernández *, Sergio Esteban-Romero, Fernando Fernández-Martínez and Manuel Gil-Martín

Grupo de Tecnología del Habla y Aprendizaje Automático (THAU Group), Information Processing and Telecommunications Center, E.T.S.I. de Telecomunicación, Universidad Politécnica de Madrid (UPM), 28040 Madrid, Spain; sergio.estebanro@upm.es (S.E.-R.); fernando.fernandezm@upm.es (F.F.-M.); manuel.gilmartin@upm.es (M.G.-M.)

* Correspondence: ivan.martinf@upm.es; Tel.: +34-910672033

Abstract: The accurate modelling of video memorability, or the intrinsic properties that render a piece of audiovisual content more likely to be remembered, will facilitate the development of automatic systems that are more efficient in retrieving, classifying and generating impactful media. Recent studies have indicated a strong correlation between the visual semantics of video and its memorability. This underscores the importance of developing advanced visual comprehension abilities to enhance model performance. It has been demonstrated that Large Vision–Language Models (LVLMs) demonstrate exceptional proficiency in generalist, high-level semantic comprehension of images and video, due to their extensive multimodal pre-training on a vast scale. This work makes use of the vast generalist knowledge of LVLMs and explores efficient adaptation techniques with a view to utilising them as memorability predictors. In particular, the Quantized Low-Rank Adaptation (QLoRA) technique is employed to fine-tune the Qwen-VL model with memorability-related data extracted from the Memento10k dataset. In light of existing research, we propose a particular methodology that transforms Qwen-VL from a language model to a memorability score regressor. Furthermore, we consider the influence of selecting appropriate LoRA hyperparameters, a design aspect that has been insufficiently studied. We validate the LoRA rank and alpha hyperparameters using 5-Fold Cross-Validation and evaluate our best configuration on the official testing portion of the Memento10k dataset, obtaining a state-of-the-art Spearman Rank Correlation Coefficient (SRCC) of 0.744. Consequently, this work represents a significant advancement in modelling video memorability through high-level semantic understanding.

Keywords: large visual language models; video memorability; multimedia perception; efficient adaptation

1. Introduction

The overwhelming abundance of multimedia content presents a significant challenge in identifying which pieces will leave a lasting impression on audiences. Effective audience engagement is crucial in various sectors, including cinema, television, advertising, education, and Internet content development. Consequently, there is a critical need for systems capable of ranking, classifying, and retrieving media items based on their relevance to viewers. In this context, the concept of memorability emerges as a valuable framework for understanding and predicting the impact of multimedia content. Being able to predict how memorable a piece of audiovisual content is can help build systems that recommend, filter out, or even modify pieces of media based on this variable. Currently, systems based on

predicting media memorability are being used in the marketing field to aid in predicting the impact of marketing campaigns [1–3]. These uses can be extended to other disciplines, such as helping educators to create more memorable support materials.

Memorability is a concept rooted in human perception and is studied in the disciplines of philosophy, psychology, and neuroscience. Perception involves gathering, organizing, and processing information from the environment and is essential in various contexts, including object recognition [4] and aesthetic experiences [5]. It relies on the ability of the sensory system to use multimodal signals—concurrent stimuli in multiple sensory modalities—and is influenced by cognitive functions such as expectancy, learning, and attention [4]. Memory, a crucial cognitive function, encodes, stores, and retrieves perceived information [6]. Identifying the features of stimuli that enhance their retention in memory allows us to understand the basic mechanisms of the perceptual system and to develop artificial systems with similar capabilities.

Memorability can be studied as the innate quality of an image or video, determined by its audiovisual characteristics, that influences how easily it can be recalled in the future [7]. Perception is significantly affected by this feature of visual stimuli. Although memory may seem like a subjective cognitive process, research in psychology and neuroscience suggests that certain visual elements are inherently more memorable [8–10]. Studies have found a strong correlation between the contextual and semantic qualities of an image and its recall value [11]. Furthermore, the analysis of image semantics involves not only visual data, but also textual methods, such as descriptions of natural language [12], which highlight the multifaceted nature of memorability.

The intrinsic relationship between semantics and memorability motivates the use of tools and models that possess advanced analytical and reasoning capabilities. This could enable a deeper understanding of the complex aspects of multimedia content and how they relate to human perception. Multi-Modal Large Language Models (MM-LLMs) and Large Vision–Language Models (LVLMs) have proven themselves as the next frontier in automatic reasoning and content understanding, showcasing outstanding performance in several intricate tasks such as Visual Question–Answering [13,14], Image Captioning [15,16], and Refer Expression Comprehension [17]. Moreover, they have become powerful companion tools for instruction-based tasks, and their use as a commercial product is continuously expanding. This paper aims to leverage and adapt the advanced knowledge that LLMs present in order to generate more informed and precise memorability estimators of multimedia content. Specifically, we introduce a series of modifications performed on the original LVLMs, which are trained to generate textual output, such as descriptions or conversations, in order to turn them into memorability score regression models. We also report state-of-the-art performance in predicting memorability scores on the Memento10k corpus test set [18] with a Spearman Rank Correlation Coefficient (SRCC) of 0.744 by applying these techniques to the Qwen-VL model [19]. Our main contributions include the following:

1. Introducing the label pre-processing and prediction post-processing necessary for using LVLMs for regression.
2. Exploring Parameter-Efficient Fine-Tuning (PEFT) of LVLMs, and more concretely Quantized Low-Rank Adaptation (QLoRA) [20] of LVLMs for memorability prediction.
3. Evaluating our proposal on the Memento10k dataset, obtaining state-of-the-art performance.

The remainder of this paper is structured as follows. Section 2 provides an overview of related work, including traditional approaches to computational memorability prediction, the evolution and applications of Large Vision–Language Models (LVLMs), and their use in memorability prediction. Section 3 outlines the materials and methods used, detail-

ing our proposal, the dataset used, and the experimental setup. Section 4 presents the results and discussion, covering the outcomes of zero-shot inference, an exploration of LoRA hyperparameters, error analysis, and a comparison against state-of-the-art methods. Finally, Section 5 concludes the paper with a summary of the findings and suggestions for future work.

2. Related Work

2.1. Traditional Approaches to Computational Memorability Prediction

The processes involved in the acquisition, storage, and retrieval of visual information are not unfamiliar to psychologists and neuroscientists. Under appropriate conditions, humans tend to show an almost limitless ability to correctly recognize stimuli that are presented to them [21]. Human memory capacity has been shown to be higher for pictorial material than for other sources such as verbal [22]. Researchers have even identified activations in different areas of the brain responsible for image and object encoding, as well as their relationship with image memorability [23,24]. Interestingly, there is evidence that certain types of stimuli are inherently easier to remember than others, despite the subjective preferences and biases of the individuals under study [8–10]. Based on this, a line of research arises that tries to analyze this intrinsic characteristic of multimedia and create computational systems that try to predict memorability based on them.

In the realm of still images, seminal work by Isola et al. [7] showed that, in the absence of familiar components, certain objects and scenes tend to be more memorable than others. This pinpoints the importance of object understanding and scene parsing when predicting memorability. Perera et al. [25] built on these observations to build a system that matches human performance in the LaMem dataset [26].

A natural extension from image to video arises thanks to studies that are carried out mainly using the Memento10k [18] and VideoMem datasets [27], with special mention to the MediaEval Predicting Video Memorability Challenge that each year fosters new and bold ideas that explore innovative memorability prediction systems [28–30]. Constantin and Ionescu [31] trained a Video Transformer using the portion of the video that is, on average, the most remembered by the annotators, thus obtaining a top SRCC score of 0.665 on the official test partition of the Memento10k dataset.

Semantic properties have also been proven to be key in the analogous task of video memorability prediction. In particular, the use of textual information to convey semantic value to image descriptors has given profitable results [12,32–34]. Dumont et al. achieved superior performance by integrating low-level descriptors, scene parsing information, and event understanding together with a contextual module that took into account the similarity between each sample and the rest of the corpus [35]. Li et al. proposed an adaptive ensemble of similar modules in order to account for possible imbalances in the importance that each of them has in predicting video memorability [36]. However, in a preliminary work, Kumar et. al showed that training an architecture that simply attends to several frames in the video using a Transformer model mimics human gaze with learnt attention patterns, achieving a top SRCC score of 0.71 over the official testing set of Memento10k [37]. In conclusion, according to recent work, an accurate video memorability prediction system should combine intrinsic image properties with semantic attribute, scene, action, and object representation.

2.2. LVLMs and Their Applications

The advent of Large Language Models (LLMs) in recent times has shaken the field of artificial intelligence (AI), shifting towards a completely new paradigm. These new models have opened the door to discussions about machine consciousness, spontaneous

emotion generation, and even Artificial General Intelligence (AGI), which has seemed beyond imaginable to date [38–41]. In general, an LLM is defined as an auto-regressive, Transformer-based [42] (encoder–decoder, encoder-only, or decoder-only) model that has been pre-trained in an unsupervised manner using a vast amount of unlabeled text, usually scraped from the Web. The pre-training task is typically Next Token Prediction (NTP), where the model tries to predict the next textual element of a sequence conditioned on the previous information:

$$\mathcal{L} = -\frac{1}{T}\sum_{t=1}^{T} \log P(y_t \mid y_{<t}; \theta) \tag{1}$$

where

- T is the sequence length;
- y_t is the actual token at time t;
- $y_{<t}$ represents all tokens before time t;
- $P(y_t \mid y_{<t}; \theta)$ is the probability of the token y_t given the preceding tokens $y_{<t}$ and model parameters θ.

State-of-the-art Large Language Models (LLMs) have demonstrated exceptional performance in a wide range of Natural Language Processing (NLP) tasks, including machine translation, document understanding, code generation, and mathematical reasoning [43–47]. Although these cutting-edge models prioritize performance, they often require substantial computational resources. In contrast, more efficient LLMs, which maintain a strong subset of these capabilities, have shown remarkable adaptability and effectiveness in handling complex tasks [48–50].

Recent advancements in artificial intelligence have shifted from text-only Large Language Models (LLMs) to Large Vision–Language Models (LVLMs) and, more broadly, Multi-Modal LLMs (MM-LLMs). These models aim to extend their inherent general knowledge to comprehend and process additional modalities, such as images, videos, and audio, either independently or in combination.

LVLMs typically integrate a pre-trained image encoder with a "translation module", which maps visual embeddings into the latent input space of the LLM. This adaptation process often employs multi-stage training on extensive multimodal datasets [19,51–53]. Leveraging their robust pre-training, advanced reasoning capabilities, and adaptability, LVLMs have demonstrated significant efficacy across a broad range of computer vision tasks [54,55]. However, a key challenge remains in refining LVLMs from general-purpose conversational agents into task-specific systems optimized for classification or regression.

2.3. LVLMs for Memorability Prediction

Cutting-edge approaches for memorability prediction have relied on LVLMs as part of their solution because of their aforementioned advantages. In recent preliminary studies, Singh et al. [56] reported that the instruction fine-tuning of the LLaMa-VID model with behavior content in the form of YouTube comments, views, likes, etc., improves its content-understanding abilities, including memorability prediction, with a reported top SRCC score of 0.71 over the validation set of the Memento10k corpus. SI et al. [57] proposed a system to predict long-term ad memorability based on end-to-end adaptation of a Llama LLM [50] and an EVA-CLIP image encoder [58]. They reported a top SRCC result of 0.75 over the validation set of the Memento10k corpus. Both approaches yielded promising results by leveraging the capabilities of LVLM-based architectures. However, they require training billions of parameters, which requires significant computational resources and time. In this work, we explore a method for adapting these large models to achieve comparable

performance while training a reduced set of parameters, which reduces computational cost and training time. By applying the QLoRA technique to the Qwen-VL LVLM, we aim to efficiently leverage the vast knowledge present in the original system, obtained through large-scale pretraining, shifting its behavior to a memorability prediction and achieving cutting-edge performance without the computational demands required for full fine-tuning. While there exists evidence of the wide capabilities of LVLMs in complex multimodal reasoning and semantic understanding, applying these skills to a downstream task with limited resources poses a challenge that we tackle through efficient fine-tuning.

3. Materials and Methods

In this section, we detail our proposed solution for LVLM-based memorability prediction and the setup employed to carry out the necessary experiments. Section 3.1 will explain our proposed architecture, including the base LVLM, the prompting technique, and the adaptation module. We will introduce the data used for experimentation in Section 3.2 and the experimental setup in Section 3.3.

3.1. Proposal

A schematic view of our proposal is shown in Figure 1. Our main objective is to benefit from the extensive and general pre-training knowledge displayed by LVLMs, steering their behavior from open-world language generation to memorability score prediction. We employ the Qwen-VL model [19] due to its relatively small size (9.6 billion parameters total), versatility, and robustness. Qwen-VL is built upon the Qwen-7B language model, enhanced with a visual processing capability. Its architecture comprises three key components: a visual encoder, a position-aware vision–language adapter, and an LLM as a text decoder. The visual encoder is based on a Vision Transformer (ViT), which processes up to three input images by segmenting them into patches and generating a sequence of image features. To efficiently integrate visual and textual modalities, the model incorporates a position-aware vision–language adapter, which employs a single-layer cross-attention module with learnable query embeddings. This adapter compresses the visual feature sequence to a fixed length of 256 while preserving spatial information through 2D absolute positional encodings. The transformed visual representations are then fed into the Qwen-7B LLM, which is responsible for generating responses based on both visual and textual inputs. The input–output interface utilizes special tokens to delineate different modalities, effectively enabling attentive processing between visual and textual tokens. The model undergoes a three-stage training process: large-scale pretraining on image–text pairs, multi-task fine-tuning with high-resolution images, and supervised instruction tuning for interactive vision–language understanding.

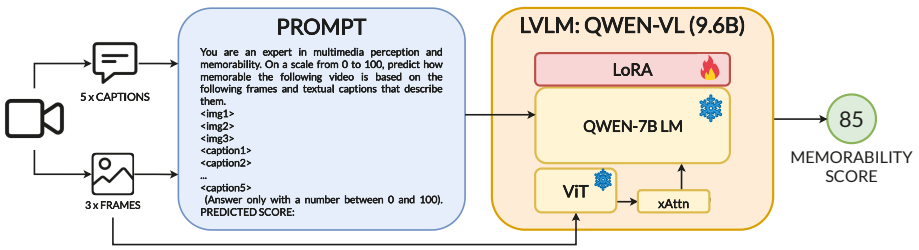


Figure 1. A visual representation of our proposed framework for predicting video memorability. Given three video frames and five textual captions as input, our model, powered by the Qwen-VL LVLM, estimates the likelihood of the video being remembered as a score ranging from 0 to 100. The weights of the LVLM are kept frozen, as indicated by the snowflake, and the Low-Rank Adaptation module is trained on the downstream task (flame symbol).

In this work, we use the `Qwen-VL-Chat-Int-4` checkpoint (https://huggingface.co/Qwen/Qwen-VL-Chat-Int4 (accessed on 4 March 2025), which incorporates two main advantages over the base model. Firstly, it has undergone Supervised Fine-Tuning (SFT) adaptation, making it more versatile and flexible. Secondly, its weights have been quantized to 4-bit integers, which drastically reduces size whilst keeping most of its predictive abilities. This goes hand in hand with our goal of benefiting from large-scale models with reduced computing resources.

The prompt used to send visual and textual information about the video and task to the model is shown in Listing 1. The first two sentences are used to instruct the model and condition its generation. Then, we include three frames extracted from the video: the first, middle, and last. Given that the videos in Memento10k are short and recorded in amateur setups, they often include a single shot and semantic unit. Therefore, this frame selection strategy has been used in previous work as a compromise between the overall representation of the visual information in the scene and computational resources [32,34,59–62]. In our case, the Qwen-VL-Chat model under study accepts a maximum of three images, which justifies our choice of this method. Finally, we append the five available textual descriptions for that video and conclude the prompt with a last conditioning sentence. The model is prompted to generate an integer score between 0 and 100 to control variance originating from the generation of uncontrolled non-integer numbers, as has been performed in a previous work [57].

Listing 1. Prompt used to feed the Qwen-VL model with visual and textual information about the video as well as context about the task and a call for action.

```
You are an expert in multimedia perception and memorability. On a
    scale from 0 to 100, predict how memorable the following
   video is based on the following frames and textual captions
   that describe them.
<img1>
<img2>
<img3>
<caption1>
<caption2>
...
<caption5>
 (Answer only with a number between 0 and 100). PREDICTED SCORE:
```

The images are encoded by a Visual Transformer (ViT) and then projected into the language embedding space via cross-attention with a set of learnable query vectors before being fed to the language model. To adapt the model for memorability prediction, we use Low-Rank Adaptation (LoRA) [20], a PEFT technique that freezes the pre-trained LVLM parameters and adds a small, trainable network to learn a linear transformation of the output. This allows efficient specialization of a large, general-purpose model, requiring significantly less computational effort than fine-tuning the full model. The adaptation process can be defined as follows:

$$h = W_0 x + \frac{\alpha}{r} \Delta W x \tag{2}$$

where

- h is the output of the adapted network;
- x is the input;
- W_0 is the weight matrix of the original network;

ΔW represents the learnt LoRA weight matrix of rank r;

α is a multiplicative factor that controls the importance given to the appended weights versus the original.

The hyperparameters α and r control the influence of the newly adapted weights. Although the authors of the original formulation set them empirically and suggest a ratio of $\alpha/r = 2$, recent works have shown that the choice of these values is highly problem-dependent and can have a significant impact on downstream performance [63,64]. In this case, where model weights are casted to 4-bit integers, the adaptation process is referred to as Quantized LoRA or QLoRA [65].

3.2. Dataset

We carry out our experiments using the Memento10k dataset [18], a collection of 10,000 videos that last 3 seconds on average. The scores are continuous: a number between 0 and 1 is associated with the ability of certain video to be remembered in general. These annotations are generated through a memory game, wherein participants watch a series of videos, which may include repetitions, and indicate by pressing a button whether they remember viewing them earlier during the game. The objective is to assess how video recall diminishes as the interval between repetitions extends in the previously mentioned memory game. Considering the brevity of the videos, it is likely that they embody a singular semantic concept.

A notable characteristic of the dataset is that the videos originated from web scraping, which generally results in reduced image quality due to recording with less expensive consumer devices. However, the emphasis is on human actions and movements, with the samples exhibiting substantial optical flow. The Memento10k dataset includes five textual descriptions per video, referred to as closed captions (CC), or simply captions. These short sentences concisely encapsulate the main attributes of each clip, emphasizing semantic content while excluding emotional elements. This characteristic favors the integration of semantic information through natural languages, thus enriching the original visual attributes.

3.3. Experimental Setup

Regarding our internal experiments, we evaluate the effect of QLoRA fine-tuning against zero-shot inference using the original Qwen-VL model. We also explore the impact of the LoRA hyperparameters on downstream performance. To that end, we evaluate our experiments using 5-Fold Cross-Validation using the concatenation of all labeled samples from the Memento10k corpus—training and validation sets, comprising a grand total of 8500 videos. In this way, we isolate the performance evaluation from the possible particularities of the validation set distribution. We use the mean Spearman Correlation Coefficient over the five folds as a figure of merit, in concordance with the literature. Lastly, we perform inference on the 1500 unlabeled samples from the test set that is used as part of the MediaEval Predicting Video Memorability challenge [28–30]. We submit our run to the challenge organizers in order to obtain a performance metric on said data. All models are trained with a batch size of 4 for 4 epochs using a single 40 GB NVIDIA A100 GPU (NVIDIA, Santa Clara, CA, USA).

4. Results and Discussion

4.1. Zero-Shot Inference

First, we explore the zero-shot capabilities of Qwen-VL by performing inference on all labeled data from Memento10k using the prompt described in Section 3.1. As can be seen

in Figure 2, when used in inference mode the LVLM might not always adjust to the desired numerical output but rather produce full sentences containing a possibly valid numerical value that must be extracted. For that reason, we post-process the results, using regular expressions to find and extract a number between 0 and 100 from the model output. If no such number is found, we consider the prediction invalid and assign the mean score of the entire training dataset (80) for evaluation. After filtering 600 invalid responses from the 8500 total (7.06%), we obtain an SRCC value of $-\mathbf{0.050}$. This suggests that, in spite of the large amount of pre-training data and the vast knowledge acquired in the first learning stages, memorability prediction as a regression problem falls out of the scope of the zero-shot capabilities of the model. This intuition is supported by prediction points, as shown in Figure 3. The predicted values are shown to be binned to a few discrete values that show no correlation with the ground truth. Therefore, domain adaptation is necessary to guide the model in relating visual and semantic cues from the video samples effectively to their potential memorability.

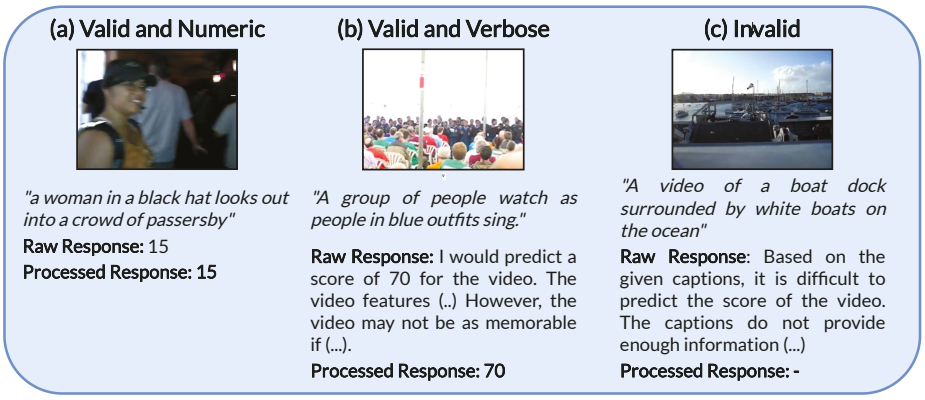


Figure 2. Examples of responses given by the zero-shot model when prompted about memorability scores. The answer can be just a number (**a**), a paragraph including a numerical result that can be correctly parsed (**b**), or text with no valid integer prediction (**c**).

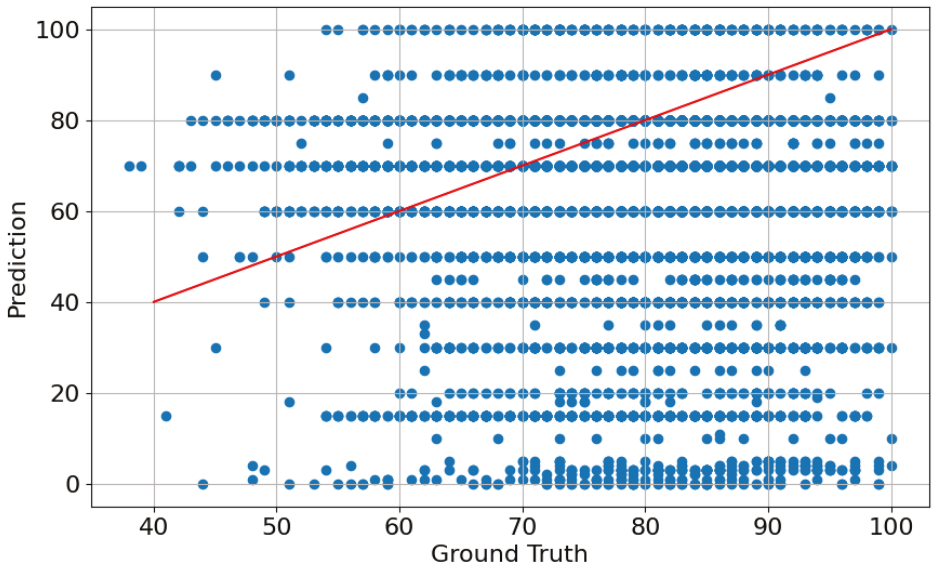


Figure 3. Scatter plot of the validation predictions and ground truths for the zero-shot Qwen-VL model. Each blue dot represents a video sample. The red line represents the perfect estimator: Ground Truth = Prediction.

4.2. LoRA Hyperparameter Exploration

The mean performance in terms of SRCC in the five folds for all alpha-rank configurations tested is shown in Figure 4. First, it should be noted that all results exceed 0.73, demonstrating the effectiveness of this method. By training a relatively smaller network, the pre-trained knowledge of the base LVLM can be leveraged to build a strong memorability predictor. Moreover, a top result of **0.7553** is obtained using 128 rank matrices with an alpha parameter of 1024. In general, the observed tendencies highlight the significant impact of selecting an appropriate alpha parameter. A low value suggests under-utilization of the adaptation module, therefore neglecting the acquired task-specific knowledge. Conversely, an excessively high alpha value can diminish performance by overshadowing the base inherent image and language understanding capabilities held by the model. Overall, these results demonstrate the beneficial synergy between a large general-purpose foundation model and a smaller task-specific adapter.

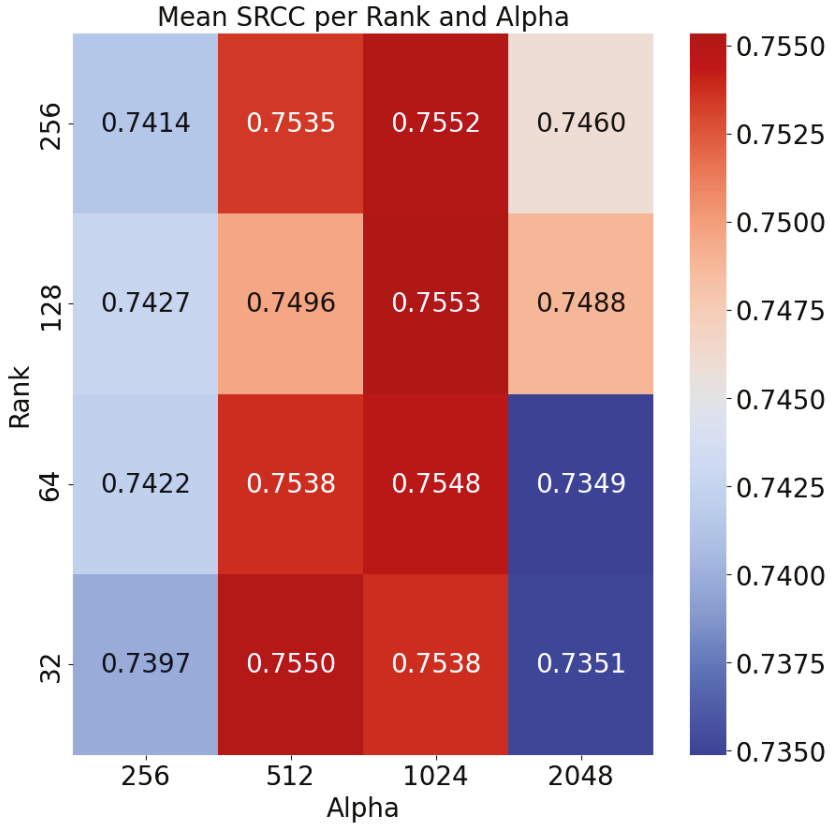


Figure 4. Heatmap of the mean SRCC over the five folds for each alpha-rank configuration tested.

With respect to the *rank* parameter, it appears to have a less pronounced effect on model performance. This value represents the rank of the matrix ΔW that is learned as an adapter. Although performance variations are not substantial, achieving the best result with a rank of 128 suggests that, for this particular problem and model setup, a rank-128 decomposition most effectively characterizes the original weight matrices. Overall, this study demonstrates that the appropriate selection of LoRA hyperparameters can greatly impact the performance of the model. Although we used a grid search for this purpose, future work should explore more efficient hyperparameter optimization strategies.

4.3. Error Analysis

Next, we explore the predictions obtained using the best model ($r = 128$, $\alpha = 1024$). Figure 5 shows a scatter plot for the predicted and ground truth scores obtained using this

checkpoint. It can be seen that, for most of the range, there is a clear correlation between both, in line with the obtained metric. However, it can also be noticed that the model tends to predict better samples with higher memorability, whilst it tends to assign higher values to the bottom-tier videos. In particular, there are no predicted scores below 50, whilst the ground truth values span below 40. This aspect might be derived from the fact that there are relatively few low-memorability examples in the dataset, which prevents the model from learning generalizable characteristics of videos that are associated with a lower probability of being remembered, or even including these lower values in its range of predictions. Moreover, as can be seen in Figure 6, the worst predicted examples are also in the low-memorability zone, supporting the hypothesis that the model lacks understanding of what makes a video forgettable. Visual inspection shows that the videos that are best predicted often show distinct objects or people performing clear actions such as playing, singing, or facing the camera in close-up shots. On the other hand, the worst-performing samples (which also tend to be lower-valued) often depict poorly lit scenarios or lack of a distinctive protagonist figure, let it be an object, animal, or person. This underperformance can be attributed to fundamental challenges in visual processing that directly impact memorability. Poor lighting reduces available visual information, leading to a lower signal-to-noise ratio, loss of detail, and color distortion. These degradations hinder the effectiveness of feature extraction by image encoders such as ViTs, creating a domain shift from well-lit training data. Furthermore, the absence of a distinct protagonist introduces ambiguity, making object and action recognition, as well as general scene understanding, more difficult. This lack of a clear focal point disrupts attention mechanisms, leading to less informative representations. Critically, these visual deficiencies directly impact memorability, as poorly lit and ambiguous scenes lacking a salient focus are inherently less memorable. This aligns with human perception research, which emphasizes the role of lighting, object salience, and scene complexity in memory encoding.

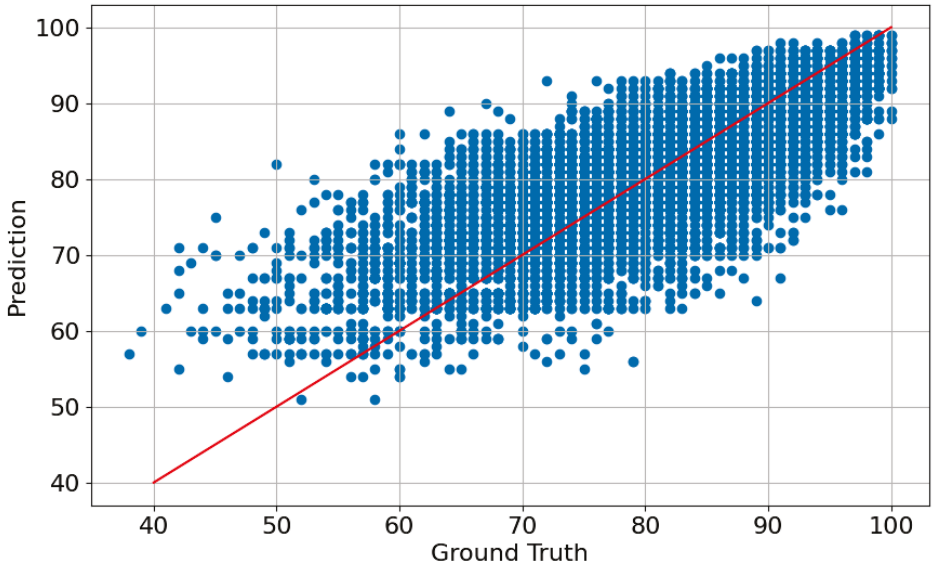


Figure 5. Scatter plot of the validation predictions and ground truths for the best-performing model on the five folds ($r = 128$, $\alpha = 1024$). Each blue dot represents a video sample. The red line represents the perfect estimator: Ground Truth = Prediction.

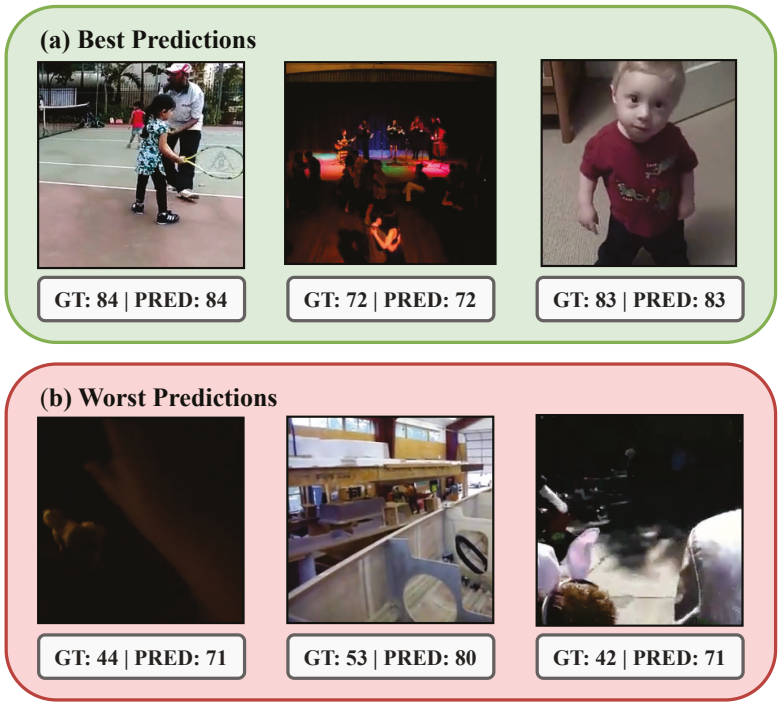


Figure 6. Best (**a**) and worst (**b**) memorability predicted examples from the validation set. GT: Ground Truth, PRED: Prediction.

4.4. Comparison Against the State of the Art

After choosing the optimum hyperparameters, we train a new checkpoint with all available labeled data and perform inference on the unlabeled test samples, submitting a run to the MediaEval challenge organizers for a fair test evaluation on unseen data. Our system achieves an SRCC of **0.7444**, which notably surpasses previous peer-reviewed works under the same setup (see Table 1). This serves as evidence of the robustness of our method that even exceeds human consistency. It also highlights the adequate choice of experimental setup. By performing a Cross-Validation hyperparameter selection instead of evaluating over a single validation partition, we improve the robustness of our estimation of system performance. This is evidenced by a drop in performance from the validation results to the test results of 1.46%, which is relatively low for this problem and configuration. Previous state-of-the-art solutions relied on incorporating several visual and textual features through specialist modules [18,35] or training large vision backbones with a high amount of visual data per sample [31,37]. On the other hand, our approach leverages both textual and visual pre-training knowledge through PEFT and low-resource adaptation, using only three frames and five textual descriptions per video. In this way, we put this large-scale knowledge into service as a powerful video memorability prediction. Moreover, the QLoRA method enables this LVLM to be adapted for a downstream task with proficient results by training a limited fraction of the total parameters. For the best-performing configuration ($r = 128 \quad \alpha = 1024$), out of the total 9.6 B model weights, only 67.1 M (0.7%) are updated. Although the original Qwen-VL model has a greater total number of parameters in comparison with previous state-of-the-art models, these strategies rely on either training from scratch or fine-tuning the whole network, failing to effectively leverage base knowledge acquired by foundation models. Our approach excels in applying minimal transformations to the base checkpoints in an effective manner, allowing model size to be scaled without increasing computational requirements by the same order of magnitude.

Together with quantization, low-rank adaptation enables both training and inference of this otherwise expensive model on a single GPU.

Table 1. Comparison against the state of the art for our proposed solution. The results shown are reported over the official test partition of the Memento10k dataset. The number of parameters for all the approaches except ours are estimated from the available information in the published papers and related code releases. For Kumar et al. [37], the available information lacks fundamental structural details (i.e., number of layers and heads of the Transformer model) to provide a reasonable estimation. The best result in terms of SRCC is highlighted in bold. N/A stands for Not Available.

Model	Number of Total Parameters	Number of Trained Parameters (% Total)	SRCC (95% CI)
Human Consistency [18]	-	-	0.730
SemanticMemNet [18]	13 M	13 M (100%)	0.663 (0.634, 0.690)
Constantin, Ionescu [31]	57.7 M	57.7 M (100%)	0.665 (0.636, 0.692)
M3-S [35]	110 M	110 M (100%)	0.670 (0.641, 0.697)
Kumar et al. [1] [37]	N/A	N/A	0.713 (0.687, 0.737)
Qwen-VL (QLoRA) (Ours)	9.6 B	67.1 M (0.7%)	**0.744 (0.721, 0.766)**

[1] Preprint, accepted for publication at WACV 2025.

4.5. Limitations

While our method demonstrates strong performance on the official test portion of Memento10k, the reference dataset in academic memorability prediction, several limitations remain. Memento10k is a highly specific dataset, featuring short, low-quality videos that typically depict a single semantic unit. This constrained setting raises questions about the generalizability to broader video types, such as movie excerpts, commercial advertisements, or educational multimedia. In such cases, the assumption that three frames can sufficiently represent an entire video may not hold, necessitating the use of LVLMs capable of processing full video sequences. Future work should explore whether the QLoRA adaptation strategy generalizes to such models and scenarios.

Additionally, while QLoRA fine-tuning enables efficient adaptation, our approach leverages a large-scale LVLM, Qwen-VL, which has a significantly higher parameter count than previous state-of-the-art models. Further research should investigate alternative video-processing LVLMs with comparable or smaller total parameter counts, balancing efficiency and performance.

5. Conclusions and Future Work

This study demonstrates the efficacy of adapting Large Vision–Language Models (LVLMs) for video memorability prediction, a task that takes advantage of the intrinsic relationship between semantic understanding and human perception. Using the QLoRA framework for parameter-efficient fine-tuning, we successfully transformed the Qwen-VL model into a state-of-the-art memorability predictor, achieving a Spearman Rank Correlation Coefficient (SRCC) of 0.744 on the Memento10k official test partition dataset. This significant improvement over existing methods underscores the potential to leverage generalist multimodal pre-training in specialized domains. Notably, our results demonstrate that although parameter-efficient methods maintain a significant portion of the underlying knowledge of the pre-trained model, they exhibit considerable sensitivity to task-specific adjustments.

Future work will focus on addressing the aforementioned limitations and expanding the scope of research in several directions. First, we aim to incorporate additional datasets to diversify the range of content and enhance the generalizability of the model. Second, alternative parameter-efficient adaptation methods, such as prompt-tuning and adapter-based learning, could be explored to further optimize performance. Third, integrating more sophisticated semantic representations, such as contextual embeddings from natural language descriptions or affective elements, may provide deeper insights into the cognitive mechanisms underlying memorability. The assimilation of additional modalities and more nuanced information is motivated by recent advances in rich, semantical audio–visual processing and video and scene understanding through efficient memory handling on LLMs, amongst others [66–68]. Finally, an investigation of fairness and bias in model predictions will be critical to ensure equitable outcomes in various types of content and audiences.

Author Contributions: Conceptualization, I.M.-F., F.F.-M. and M.G.-M.; methodology, I.M.-F., F.F.-M. and M.G.-M.; software, I.M.-F. and S.E.-R.; investigation, I.M.-F. and S.E.-R.; resources, F.F.-M. and M.G.-M.; writing—original draft preparation, I.M.-F.; writing—review and editing, I.M.-F., S.E.-R., F.F.-M. and M.G.-M.; supervision, F.F.-M. and M.G.-M.; project administration, F.F.-M. and M.G.-M.; funding acquisition, F.F.-M. and M.G.-M. All authors have read and agreed to the published version of the manuscript.

Funding: The research of Iván Martín-Fernández was supported by the Universidad Politécnica de Madrid (Programa Propio I+D+i). Sergio Esteban-Romero's research was supported by the Spanish Ministry of Education (FPI grant PRE2022-105516). This work was funded by Project ASTOUND (101071191—HORIZON-EIC-2021-PATHFINDERCHALLENGES-01) of the European Commission and by the Spanish Ministry of Science and Innovation through the projects GOMINOLA (PID2020-118112RB-C22), TRUSTBOOST (PID2023-150584OB-C21) and BeWord (PID2021-126061OB-C43), funded by MCIN/AEI/10.13039/501100011033 and by the European Union "NextGenerationEU/PRTR".

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Restrictions apply to the availability of these data. Data were obtained from Newman et al. [18] and are available from the authors with their permission.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Memorable.io. Memorable.io: AI-Powered Memorability Prediction. 2025. Available online: https://www.memorable.io/ (accessed on 4 March 2025).
2. Neosperience. Image Memorability by Neosperience. 2025. Available online: https://image.neosperience.com/ (accessed on 4 March 2025).
3. Needell, C.D. Resmem: A Package that Wraps the ResMem Pytorch Model. 2021. Available online: https://github.com/BrainBridge-Lab/resmem (accessed on 4 March 2025).
4. Bernstein, D.A.; Nash, P.W. Sensation and Perception. In *Essentials of Psychology*; Houghton Mifflin Company: Boston, MA, USA, 2008; pp. 84–134.
5. Leder, H.; Belke, B.; Oeberst, A.; Augustin, D. A model of aesthetic appreciation and aesthetic judgments. *Br. J. Psychol.* **2004**, *95*, 489–508. [CrossRef] [PubMed]
6. Bernstein, D.A.; Nash, P.W. Memory. In *Essentials of Psychology*; Houghton Mifflin Company: Boston, MA, USA, 2008; pp. 207–245.
7. Isola, P.; Xiao, J.; Torralba, A.; Oliva, A. What makes an image memorable? In Proceedings of the CVPR 2011 IEEE, Colorado Springs, CO, USA, 20–25 June 2011; pp. 145–152.
8. Isola, P.; Parikh, D.; Torralba, A.; Oliva, A. Understanding the intrinsic memorability of images. *Adv. Neural Inf. Process. Syst.* **2011**, *24* , 2429–2437.

9. Lin, Q.; Yousif, S.R.; Chun, M.M.; Scholl, B.J. Visual memorability in the absence of semantic content. *Cognition* **2021**, *212*, 104714. [CrossRef] [PubMed]
10. Xie, W.; Bainbridge, W.A.; Inati, S.K.; Baker, C.I.; Zaghloul, K.A. Memorability of words in arbitrary verbal associations modulates memory retrieval in the anterior temporal lobe. *Nat. Hum. Behav.* **2020**, *4*, 937–948. [CrossRef]
11. Bylinskii, Z.; Goetschalckx, L.; Newman, A.; Oliva, A. Memorability: An image-computable measure of information utility. In *Human Perception of Visual Information: Psychological and Computational Perspectives*; Springer: Cham, Switzerland, 2022; pp. 207–239.
12. Kleinlein, R.; Luna-Jiménez, C.; Arias-Cuadrado, D.; Ferreiros, J.; Fernández-Martínez, F. Topic-Oriented Text Features Can Match Visual Deep Models of Video Memorability. *Appl. Sci.* **2021**, *11*, 7406. [CrossRef]
13. Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; Parikh, D. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6904–6913.
14. Lu, P.; Mishra, S.; Xia, T.; Qiu, L.; Chang, K.W.; Zhu, S.C.; Tafjord, O.; Clark, P.; Kalyan, A. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 2507–2521.
15. Agrawal, H.; Desai, K.; Wang, Y.; Chen, X.; Jain, R.; Johnson, M.; Batra, D.; Parikh, D.; Lee, S.; Anderson, P. Nocaps: Novel object captioning at scale. In Proceedings of the IEEE/CVF international Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8948–8957.
16. Young, P.; Lai, A.; Hodosh, M.; Hockenmaier, J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguist.* **2014**, *2*, 67–78. [CrossRef]
17. Kazemzadeh, S.; Ordonez, V.; Matten, M.; Berg, T. Referitgame: Referring to objects in photographs of natural scenes. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 787–798.
18. Newman, A.; Fosco, C.; Casser, V.; Lee, A.; McNamara, B.; Oliva, A. Multimodal memorability: Modeling effects of semantics and decay on video memorability. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XVI 16; Springer: Cham, Switzerland, 2020; pp. 223–240.
19. Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; Zhou, J. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv* **2023**, arXiv:2308.12966.
20. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv* **2021**, arXiv:2106.09685v1.
21. Shepard, R.N. Recognition memory for words, sentences, and pictures. *J. Verbal Learn. Verbal Behav.* **1967**, *6*, 156–163. [CrossRef]
22. Standing, L. Learning 10,000 pictures. *Q. J. Exp. Psychol.* **1973**, *25*, 207–222.
23. Jaegle, A.; Mehrpour, V.; Mohsenzadeh, Y.; Meyer, T.; Oliva, A.; Rust, N. Population response magnitude variation in inferotemporal cortex predicts image memorability. *eLife* **2019**, *8*, e47596. [CrossRef]
24. Lahner, B.; Mohsenzadeh, Y.; Mullin, C.; Oliva, A. Visual perception of highly memorable images is mediated by a distributed network of ventral visual regions that enable a late memorability response. *PLoS Biol.* **2024**, *22*, e3002564. [CrossRef] [PubMed]
25. Perera, S.; Tal, A.; Zelnik-Manor, L. Is image memorability prediction solved? In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 15–20 June 2019.
26. Khosla, A.; Raju, A.S.; Torralba, A.; Oliva, A. Understanding and predicting image memorability at a large scale. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2390–2398.
27. Cohendet, R.; Demarty, C.H.; Duong, N.Q.; Engilberge, M. VideoMem: Constructing, analyzing, predicting short-term and long-term video memorability. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 2531–2540.
28. Constantin, M.G.; Demarty, C.H.; Fosco, C.; Seco, A.G.; de Herrera, S.H.; Healy, G.; Ionescu, B.; Matran-Fernandez, A.; Kiziltepe, R.S.; Smeaton, A.F.; et al. Overview of The MediaEval 2023 Predicting Video Memorability Task. In Proceedings of the MediaEval '23: Multimedia Evaluation Workshop, Amsterdam, The Netherlands, 1–2 February 2024.
29. Sweeney, L.; Constantin, M.G.; Demarty, C.H.; Fosco, C.; de Herrera, A.G.S.; Halder, S.; Healy, G.; Ionescu, B.; Matran-Fernandez, A.; Smeaton, A.F.; et al. Overview of the MediaEval 2022 predicting video memorability task. *arXiv* **2022**, arXiv:2212.06516.
30. Kiziltepe, R.S.; Constantin, M.G.; Demarty, C.H.; Healy, G.; Fosco, C.; de Herrera, A.G.S.; Halder, S.; Ionescu, B.; Matran-Fernandez, A.; Smeaton, A.F.; et al. Overview of the MediaEval 2021 predicting media memorability task. *arXiv* **2021**, arXiv:2112.05982.
31. Constantin, M.G.; Ionescu, B. AIMultimediaLab at MediaEval 2022: Predicting Media Memorability Using Video Vision Transformers and Augmented Memorable Moments. In Proceedings of the MediaEval '23: Multimedia Evaluation Workshop, Bergen, Norway, 13–15 January 2023.

32. Martín-Fernández, I.; Kleinlein, R.; Luna-Jiménez, C.; Gil-Martín, M.; Fernández-Martínez, F. Video Memorability Prediction From Jointly-learnt Semantic and Visual Features. In Proceedings of the 20th International Conference on Content-Based Multimedia Indexing, New York, NY, USA, 20–22 September 2023; pp. 178–182. [CrossRef]
33. Agarla, M.; Celona, L.; Schettini, R.; et al. Predicting Video Memorability Using a Model Pretrained with Natural Language Supervision. In Proceedings of the MediaEval Multimedia Benchmark Workshop Working Notes, Bergen, Norway and Online, 12–13 January 2023; Volume 1.
34. Kleinlein, R.; Luna-Jiménez, C.; Fernández-Martínez, F. THAU-UPM at MediaEval 2021: From Video Semantics To Memorability Using Pretrained Transformers. In Proceedings of the MediaEval Multimedia Benchmark Workshop Working Notes, Online, 13–15 December 2021.
35. Dumont, T.; Hevia, J.S.; Fosco, C.L. Modular memorability: Tiered representations for video memorability prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 10751–10760.
36. Li, J.; Guo, X.; Yue, F.; Xue, F.; Sun, J. Adaptive Multi-Modal Ensemble Network for Video Memorability Prediction. *Appl. Sci.* **2022**, *12*, 8599. [CrossRef]
37. Kumar, P.; Khandelwal, E.; Tapaswi, M.; Sreekumar, V. Seeing Eye to AI: Comparing Human Gaze and Model Attention in Video Memorability. *arXiv* **2024**, arXiv:2311.16484.
38. Elkins, K.; Chun, J. Can GPT-3 pass a Writer's turing test? *J. Cult. Anal.* **2020**, *5* , 1–16. [CrossRef]
39. Kosinski, M. Theory of mind may have spontaneously emerged in large language models. *arXiv* **2023**, arXiv:2302.02083.
40. Yukun, Z.; Xu, L.; Huang, Z.; Peng, K.; Seligman, M.; Li, E.; Yu, F. AI Chatbot Responds to Emotional Cuing. *PsyArXiv* **2023**, preprint. . [CrossRef]
41. Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y.T.; Li, Y.; Lundberg, S.; et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv* **2023**, arXiv:2303.12712.
42. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.u.; Polosukhin, I. Attention is All you Need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Newry, UK, 2017; Volume 30.
43. Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A.M.; Hauth, A.; Millican, K.; et al. Gemini: A family of highly capable multimodal models. *arXiv* **2023**, arXiv:2312.11805.
44. Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. The llama 3 herd of models. *arXiv* **2024**, arXiv:2407.21783.
45. OpenAI. GPT-4 Technical Report. *arXiv* **2023**, arXiv:2303.08774.
46. Yang, A.; Yang, B.; Hui, B.; Zheng, B.; Yu, B.; Zhou, C.; Li, C.; Li, C.; Liu, D.; Huang, F.; et al. Qwen2 Technical Report. *arXiv* **2024**, arXiv:2407.10671.
47. Anthropic. *Claude 3.5 Sonnet Model Card Addendum*; Anthropic: San Francisco, CA, USA, 2024.
48. Team, G.; Riviere, M.; Pathak, S.; Sessa, P.G.; Hardin, C.; Bhupatiraju, S.; Hussenot, L.; Mesnard, T.; Shahriari, B.; Ramé, A.; et al. Gemma 2: Improving open language models at a practical size. *arXiv* **2024**, arXiv:2408.00118.
49. Qwen Team. Qwen2.5 Technical Report. *arXiv* **2025**, arXiv:2412.15115.
50. Meta. Llama 3.2 Model Card. 2024. Available online: https://github.com/meta-llama/llama-models/blob/main/models/llama3_2/MODEL_CARD.md (accessed on 25 November 2024).
51. Beyer, L.; Steiner, A.; Pinto, A.S.; Kolesnikov, A.; Wang, X.; Salz, D.; Neumann, M.; Alabdulmohsin, I.; Tschannen, M.; Bugliarello, E.; et al. Paligemma: A versatile 3b vlm for transfer. *arXiv* **2024**, arXiv:2407.07726.
52. Liu, H.; Li, C.; Wu, Q.; Lee, Y.J. Visual Instruction Tuning. In Proceedings of the NeurIPS, New Orleans, LA, USA, 10–16 December 2024.
53. Liu, H.; Li, C.; Li, Y.; Lee, Y.J. Improved Baselines with Visual Instruction Tuning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle WA, USA, 17–21 June 2024.
54. de Zarzà, I.; de Curtò, J.; Roig, G.; Calafate, C.T. LLM Multimodal Traffic Accident Forecasting. *Sensors* **2023**, *23*, 9225. [CrossRef]
55. Ghosh, A.; Acharya, A.; Saha, S.; Jain, V.; Chadha, A. Exploring the frontier of vision-language models: A survey of current methodologies and future directions. *arXiv* **2024**, arXiv:2404.07214.
56. Singh, S.; Si, H.; Singla, Y.K.; Baths, V.; Shah, R.R.; Chen, C.; Krishnamurthy, B. LLaVA Finds Free Lunch: Teaching Human Behavior Improves Content Understanding Abilities Of LLMs. *arXiv* **2024**, arXiv:2405.00942.
57. au2, H.S.I.; Singh, S.; Singla, Y.K.; Bhattacharyya, A.; Baths, V.; Chen, C.; Shah, R.R.; Krishnamurthy, B. Long-Term Ad Memorability: Understanding and Generating Memorable Ads. *arXiv* **2024**, arXiv:2309.00378.
58. Sun, Q.; Fang, Y.; Wu, L.; Wang, X.; Cao, Y. Eva-clip: Improved training techniques for clip at scale. *arXiv* **2023**, arXiv:2303.15389.

59. Nguyen, M.Q.; Trinh, M.H.; Bui, H.G.; Vo, K.T.; Tran, M.T.; Tran, T.P.; Nguyen, H.D. SELAB-HCMUS at MediaEval 2023: A cross-domain and subject-centric approach towards the memorability prediction task. In Proceedings of the MediaEval Multimedia Benchmark Workshop Working Notes, Bergen, Norway and Online, 12–13 January 2023.
60. Usmani, M.M.A.; Zahid, S.; Tahir, M.A. Modelling of Video Memorability using Ensemble Learning and Transformers. In Proceedings of the MediaEval'22: Multimedia Evaluation Workshop, Bergen, Norway, 13–15 January 2023.
61. Azzakhnini, S.; Ahmed, O.B.; Fernandez-Maloigne, C. Video Memorability Prediction using Deep Features and Loss-based Memorability Distribution Estimation. In Proceedings of the MediaEval'22: Multimedia Evaluation Workshop, Bergen, Norway, 13–15 January 2023.
62. Sweeney, L.; Smeaton, A.; Healy, G. Memories in the Making: Predicting Video Memorability with Encoding Phase EEG. In Proceedings of the 20th International Conference on Content-Based Multimedia Indexing, Orleans, France, 20–22 September 2023; pp. 183–187.
63. Esteban-Romero, S.; Bellver-Soler, J.; Martın-Fernández, I.; Gil-Martın, M.; D'Haro, L.F.; Fernández-Martınez, F. THAU-UPM at EmoSPeech-IberLEF2024: Efficient Adaptation of Mono-modal and Multi-modal Large Language Models for Automatic Speech Emotion Recognition. In Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), Co-Located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), CEUR-WS. org. Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), Castilla y León, Spain, 24 September 2024.
64. Esteban-Romero, S.; Martín-Fernández, I.; Gil-Martín, M.; Fernández-Martínez, F. Synthesizing Olfactory Understanding: Multimodal Language Models for Image-Text Smell Matching. 2024. Available online: https://ssrn.com/abstract=4912100 (accessed on 4 March 2025).
65. Dettmers, T.; Pagnoni, A.; Holtzman, A.; Zettlemoyer, L. QLoRA: Efficient Finetuning of Quantized LLMs. In Proceedings of the Advances in Neural Information Processing Systems, New Orleans, LA, USA, 10–16 December 2023; Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S., Eds.; Curran Associates, Inc.: Newry, UK, 2023; Volume 36, pp. 10088–10115.
66. Li, K.; Yang, Z.; Chen, L.; Yang, Y.; Xiao, J. CATR: Combinatorial-Dependence Audio-Queried Transformer for Audio-Visual Video Segmentation. In Proceedings of the 31st ACM International Conference on Multimedia, New York, NY, USA, 20–22 September 2023; pp. 1485–1494. [CrossRef]
67. Yang, Z.; Chen, G.; Li, X.; Wang, W.; Yang, Y. DoraemonGPT: Toward Understanding Dynamic Scenes with Large Language Models (Exemplified as A Video Agent). In Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria, 21–27 July 2024; Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., Berkenkamp, F., Eds.; Volume 235, pp. 55976–55997.
68. Xiong, H.; Yang, Z.; Yu, J.; Zhuge, Y.; Zhang, L.; Zhu, J.; Lu, H. Streaming Video Understanding and Multi-round Interaction with Memory-enhanced Knowledge. In Proceedings of the Thirteenth International Conference on Learning Representations, Singapore, 24–28 April 2025.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

sensors

MDPI

Article

Extending Anxiety Detection from Multimodal Wearables in Controlled Conditions to Real-World Environments

Abdulrahman Alkurdi [1], Maxine He [2], Jonathan Cerna [2], Jean Clore [3], Richard Sowers [4], Elizabeth T. Hsiao-Wecksler [1,2,5] and Manuel E. Hernandez [2,5,6,7,8,*]

[1] Department of Mechanical Science & Engineering, University of Illinois Urbana-Champaign, Urbana, IL 61801, USA; alkurdi2@illinois.edu (A.A.); ethw@illinois.edu (E.T.H.-W.)
[2] Neuroscience Program, University of Illinois Urbana-Champaign, Urbana, IL 61801, USA; maoyuan2@illinois.edu (M.H.); cerna3@illinois.edu (J.C.)
[3] Department of Psychiatry and Behavioral Medicine, University of Illinois College of Medicine at Peoria, Peoria, IL 61805, USA; jayclore@uic.edu
[4] Department of Industrial & Enterprise Systems Engineering, University of Illinois Urbana-Champaign, Urbana, IL 61801, USA; r-sowers@illinois.edu
[5] Department of Biomedical and Translational Sciences, Carle Illinois College of Medicine, University of Illinois Urbana-Champaign, Urbana, IL 61801, USA
[6] Department of Health and Kinesiology, University of Illinois Urbana-Champaign, Urbana, IL 61801, USA
[7] Department of Bioengineering, University of Illinois Urbana-Champaign, Urbana, IL 61801, USA
[8] Beckman Institute, University of Illinois Urbana-Champaign, Urbana, IL 61801, USA
* Correspondence: mhernand@illinois.edu; Tel.: +1-217-244-8971

Abstract: This study quantitatively evaluated whether and how machine learning (ML) models built by data from controlled conditions can fit real-world conditions. This study focused on feature-based models using wearable technology from real-world data collected from young adults, so as to provide insights into the models' robustness and the specific challenges posed by diverse environmental noise. Feature-based models, particularly XGBoost and Decision Trees, demonstrated considerable resilience, maintaining higher accuracy and reliability across different noise levels. This investigation included an in-depth analysis of transfer learning, highlighting its potential and limitations in adapting models developed from standard datasets, like WESAD, to complex real-world scenarios. Moreover, this study analyzed the distributed feature importance across various physiological signals, such as electrodermal activity (EDA) and electrocardiography (ECG), considering their vulnerability to environmental factors. It was found that integrating multiple physiological data types could significantly enhance model robustness. The results underscored the need for a nuanced understanding of signal contributions to model efficacy, suggesting that feature-based models showed much promise in practical applications.

Keywords: anxiety; machine learning; wearable technology; transfer learning; multimodal

1. Introduction

Anxiety disorders are among the most prevalent mental health issues worldwide, affecting millions and significantly impacting the quality of life [1–3]. While traditional methods like self-reports and clinical interviews offer insight into symptoms of anxiety, they are limited by subjectivity and intermittency [4]. The advent of wearable technology presents new possibilities for the continuous, objective, and non-invasive monitoring of anxiety symptoms in real-world settings.

Prior work has demonstrated the potential of wearable devices for anxiety detection using various physiological signals, such as electrodermal activity (EDA), heart rate vari-

ability (HRV), and accelerometer data [5–8]. However, these studies are predominantly conducted in controlled laboratory settings, which may not accurately reflect realistic conditions. While wearable devices have been used to assess anxiety in social settings [9] and predict the occurrence of anxiety [10–14], further work remains on evaluating the generalizability of pre-trained anxiety classifiers on lab data to real-world data and the benefits of noise on classifier robustness.

A recent review highlighted the use of machine learning to detect anxiety using physiological signals [15]. Building on recent work [10,14], this study extends beyond the controlled laboratory setting, delving into real-world environments to assess the practicality and effectiveness of wearable technology for monitoring anxiety. Using the open-source Wearable Stress and Affect Detection (WESAD) dataset [8], which consists of multimodal wearable sensor data in different affective states, we used a subset of data to align with the wearable hardware from our study.

Traditional machine learning (ML) models are typically trained and evaluated using the same dataset. The use of a single dataset provides increased homogeneity and consistency in data characteristics for use in establishing a benchmark of model performance. Prior work examining anxiety prediction [10–14,16–22] has typically focused on the use of self-reported measures [10,11,13,20] but has more recently focused on the use of objective physiological signals under controlled conditions [12,14,16,18,19,22]. However, few studies have examined the impact of physiological and non-physiological noise on anxiety prediction [22]. Thus, the evaluation of an ML model's effectiveness under real-world conditions may provide insights into the models' robustness and the specific challenges posed by diverse environmental noise.

Transfer learning is an approach that leverages knowledge from a pre-trained model to address a related problem with limited labeled data [23,24]. By adapting the pre-trained model to the new task, transfer learning reduces the need for extensive training data and computational resources [25,26]. It is particularly beneficial when the target task has limited labeled data, is similar to the pre-trained model's task, or when training from scratch is expensive or time-consuming [26]. Transfer learning has been successfully applied in various domains, including computer vision, natural language processing, and speech [26,27]. However, few studies have examined the robustness of anxiety detection ML models to new datasets [19], and when they have, they have primarily focused on the use of controlled conditions.

This study addresses these gaps by evaluating feature-based machine learning (ML) models under a variety of real-world conditions. Feature-based ML models leverage domain expertise by providing a basis for features to use, based on prior evidence. We placed significant emphasis on distributed feature importance across different physiological signals, considering their specific failure modes which may affect detection accuracy. For example, EDA signals can be disrupted in wet or humid environments, acceleration data may be unreliable during physical activity, and electrocardiogram (ECG) signals may be affected by conductivity issues, making them challenging to integrate into wearable devices. Blood volume pulse (BVP) signals are also susceptible to disruptions caused by gaps between the device and the skin. By investigating the robustness and reliability of these models in challenging environments, this study provides an evaluation of the most effective modalities and measures for use in real-world applications.

The overarching goal of this study was to quantitatively evaluate whether and how ML models for anxiety detection built by data from wearables in controlled conditions could be applied to real-world environments. By leveraging data from young adults that reflect a wide range of environmental conditions, we aimed to identify the most informative

features and modalities for anxiety detection and evaluate the use of open-source data to pre-train an anxiety classifier for use in novel tasks and conditions.

2. Materials and Methods

To evaluate the feasibility of using an open-source dataset for transfer learning and real-world data, we used custom Python (version 3.8) data processing and feature extraction scripts on both the WESAD dataset and real-world data (see Figure 1). Custom Python scripts are available in the following code repository: https://github.com/AbdulAlkurdi/anxietyFB (accessed on 3 February 2025). The WESAD dataset [8] was used as a basis for anxiety detection with physiological signals and the introduction of synthetic noise. The study involved 15 participants, with a gender distribution of 20% female, and the use of two wearable devices. The RespiBAN Professional (PLUX Wireless Biosignal S.A., Lisbon, Portugal) was used to record ECG, EDA, EMG, RESP, skin temperature (TEMP), and 3-axis acceleration (ACC) at 700 Hz, but was downsampled to 70 Hz for all modalities. The E4 Wristband (Empatica, Inc., Boston, MA, USA) was used to record BVP, EDA, TEMP, and ACC data at varying sampling rates. The study's protocol included baseline (53%), amusement (17%), and stress (30%) conditions. Participant self-reports after each condition provided the ground truth, incorporating scales like Positive and Negative Affect Schedule (PANAS) [28], State-Trait Anxiety Inventory (STAI) [29], and Short Stress State Questionnaire (SSSQ) [30], as previously described.

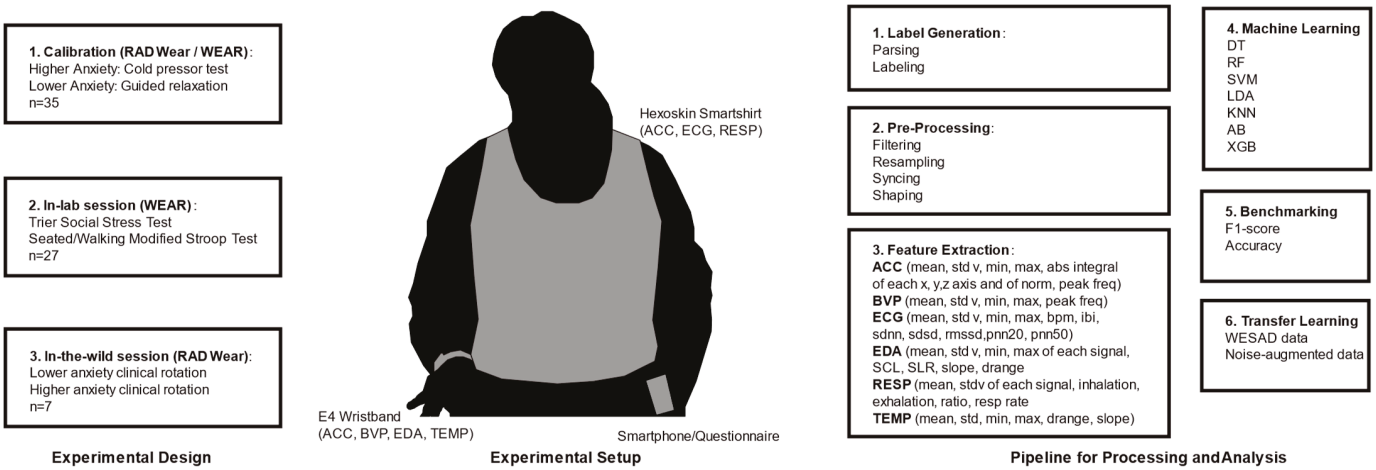


Figure 1. (**Left**) Experimental setup for in-lab and real-world data in RADWear and WEAR studies. (**Middle**) Experimental setup for data collection using wearable sensors, and (**Right**) pipeline for processing and analysis.

2.1. Overview of Real-World Data

This study was designed to quantitatively evaluate whether and how ML models for anxiety detection built by data from wearables in controlled conditions could be applied to real-world environments using data from two ongoing studies. The WEAR study captures data from STEM-major undergraduate students. Participants included 12 males and 15 females with an average height of 161 $\pm$ 5 cm, weight of 67.5 $\pm$ 20 kg, and age of 23 $\pm$ 2.3 years. The RADWear study captures the demanding nature of medical education through data from third-year students on clinical rotations. Participants included 4 males and 5 females with an average height of 167 $\pm$ 7.5 cm, weight of 79.4 $\pm$ 27 kg, and age of 27.3 $\pm$ 2.3 years. Both studies have been approved by an institutional review board at either the University of Illinois College of Medicine at Peoria (IRB Protocol 1880297 for RADWear) or the University of Illinois Urbana-Champaign (IRB Protocol 21769 for WEAR).

For both studies, the E4 wristband and Hexoskin smart shirt (Carre Technologies, Montreal, QC, Canada) were used to collect data. The E4 wristband was utilized to collect various biophysiological signals including BVP, EDA, ACC, and skin temperature (TEMP). In parallel, the Hexoskin smart shirt recorded electrocardiogram (ECG), respiration (RESP), and 3-axis acceleration (ACC) data. For the Hexoskin smartshirt, ECG data were sampled at 256 Hz, while RESP and ACC data were sampled at 128 Hz and 64 Hz, respectively.

RADWear and WEAR involved a calibration protocol designed to establish baseline states for relaxation using a guided relaxation exercise and excitement/anxiety, captured during a cold pressor test in a lab setting. These initial sessions lasted about 30 min and aimed to provide reference points for trait anxiety levels, using the State-Trait Anxiety Inventory (STAI) X-2 questionnaire at the beginning and state anxiety changes after each task, using the STAI Y6.

In the WEAR study, participants underwent a comprehensive in-lab testing session lasting approximately four hours including the calibration session. After completing the calibration session, the participants performed an in-lab session that included a series of stress- and anxiety-inducing protocols: the Trier Social Stress Test (TSST), seated Stroop test, and walking Stroop test. These activities were selected to induce anxiety and serve as validated methods to establish baseline anxiety levels. WEAR participants began with an intake survey of trait anxiety (STAI X2) and completed follow-up questionnaires of state anxiety (STAI Y6) after each test to assess anxiety responses. Prior work has demonstrated significant differences in anxiety, based on the in-person tasks, and the ability to differentiate between different tasks using objective physiological signals [31].

In the RADWear study, following the calibration protocol, participants engaged in clinical rotations, and data were collected in real-world conditions. These "in-the-wild" data collection periods were extensive, consisting of two, two-week sessions (during work hours at least 5 days per week, lasting 6–8 h each day). For the current study, nine RADWear participants were included in the calibration sessions and seven in the in-the-wild data analysis, due to availability. However, future work remains necessary to increase the sample size and validate the effectiveness of the recorded data.

The current study datasets were organized into three subsets based on the similarity of the test conditions, including physical configuration and activities involved and environmental noise (Table 1): (1) RADWear and WEAR calibration sessions, where participants are restricted to seated conditions with a fixed room temperature and lighting conditions; (2) WEAR in-lab sessions, where participants are seated, standing, or walking in a controlled environment with a fixed room temperature and lighting conditions; and (3) RADWear in-the-wild datasets, where physical configuration and activity and environmental noise are variable and unconstrained. This selection of data subsets will inform how models perform under varying levels of environmental conditions that restrict participants' motion and the potential for confounding factors.

2.2. Addressing Noise

Because of the RADWear and WEAR studies' experimental protocols containing physical activity in both lab and natural settings, feature extraction presented more of a challenge. This was due to significant amounts of motion-artifact noise, which negatively affected the signal and the ability to extract features. Given the significant impact of noise on the feature extraction of heart rate signals, a feature extraction algorithm was developed, based on concepts from Malik et al. [32] and the Automatic Multiscale-based Peak Detection (AMPD) algorithm from Scholkmann et al. [33], which was able to achieve a 98% peak detection accuracy in ECG data. For EDA and ACC signals, we used a lowpass Butterworth filter with cut-off frequencies of 5 Hz and 13 Hz, respectively. Further details

on data preprocessing and feature extraction are provided in recent work [22]. In summary, preprocessing steps included outlier removal, min–max normalization, feature extraction (see Figure 1), and handling of missing data, which were crucial for preparing RADWear and WEAR study data for effective machine learning analysis.

Table 1. Noise factors considered for each subset of the datasets.

	Physical Configuration and Activity				**Environmental Noise**	
Dataset	**Seated**	**Standing**	**Walking**	**Other**	**Fixed**	**Variable**
WESAD	■	■			■	
Calibration (RADWear/WEAR)	■				■	
In-lab session (WEAR)	■	■	■		■	
In-the-wild session (RADWear)	■	■	■	■		■

Note: Black color used to indicate physical configuration and activity or environmental noise condition in each dataset.

Feature extraction was performed on all available modalities from the E4 wristband and Hexoskin smart shirt, using features (Figure 1) that had shown anxiety detection robustness across a wide range of signal-to-noise ratios in XGB and DT models [22].

For the RADWear in-the-wild data, data loss due to excessive noise was partly resolved using imputation where sufficient data were available. If sufficient data in a segment were available, a mean imputation was employed to replace the missing points. If a significant portion of a segment was missing or too much noise existed, that segment was dropped.

2.3. Label Generation

Using a cutoff of 11 on the STAI Y6 (Range 6–24), we identified data segments to correspond to either lower or higher anxiety, to serve as ground truth for the feature-based ML models. A representative sample of STAI scores across calibration and in-lab sessions is provided in Figure 2.

2.4. Class Balancing

Generally, to have effective classification outcomes from ML algorithms, datasets should be balanced between classes. Imbalanced datasets can lead to biased model performance, misleading performance metrics, limited generalization, and robustness. Thus, the use of balancing techniques to ensure fair and accurate machine learning models is of the utmost importance. In this study, the undersampling of the dominant class was utilized to reduce the imbalance. The laboratory-controlled datasets (RADWear + WEAR calibration, WEAR in-lab) had good balance. Due to the nature of the RADWear in-the-wild dataset, the anxious and non-anxious classes were imbalanced, with 80% of the data being labeled as not anxious. Undersampling was performed on the RADWear in-the-wild dataset to match that of the calibration and in-lab segments. Undersampling was performed by reducing the number of data points collected from the dominant class, which preserves the number of points of the minority class (Table 2).

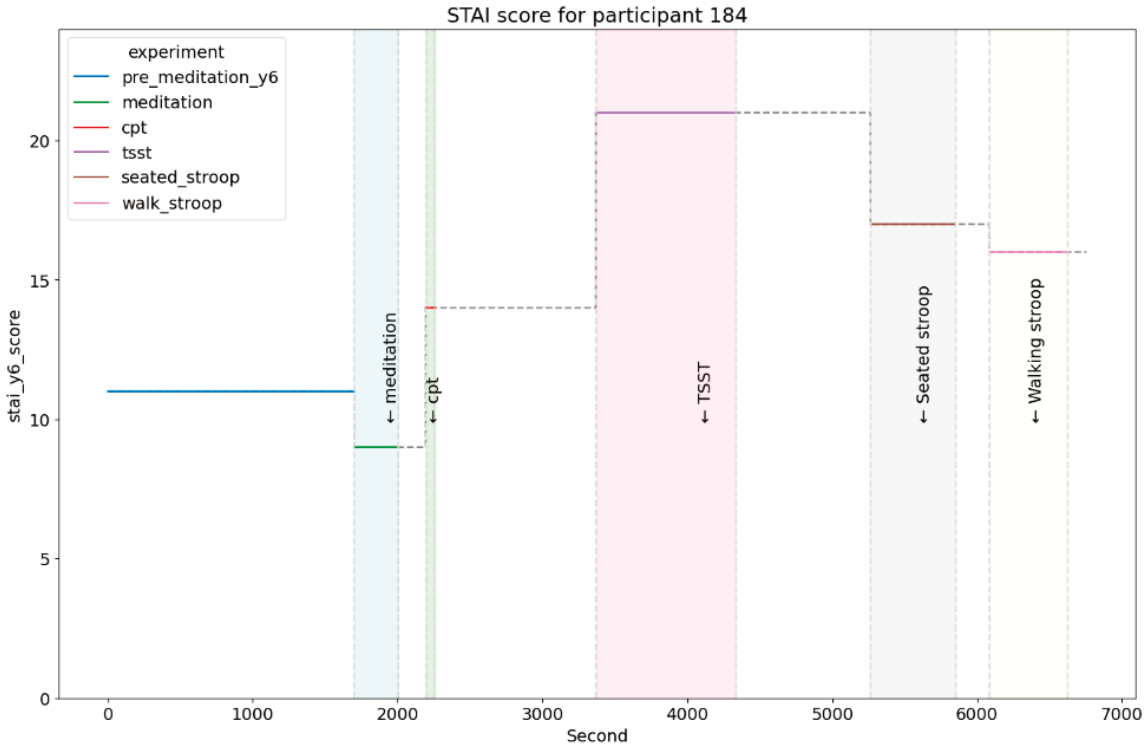


Figure 2. STAI score for one of the participants at the different test conditions experienced. The threshold considered to be indicative of the observation of anxiety is 11 for the STAI Y6. It can be observed that the meditation session reduced the anxiety level, while the cold pressor test (CPT) increased it. For this participant, the Trier Social Stress Test (TSST) seemed to be the highest cause of anxiety.

Table 2. Class distribution for each subset of the datasets.

Dataset	Not Anxious	Anxiety	Adjusted
WESAD	70%	30%	×
RADWear + WEAR Calibration	58%	42%	×
WEAR in-lab	59%	41%	×
RADWear in-the-wild	80.66%	19.34%	to 30% and 41%

2.5. Traditional Machine Learning

Based on recent findings on the most robust machine learning models in predicting anxiety from a noise-augmented WESAD dataset [22], feature-based models were employed. Feature-based models included the following: Decision Tree (DT), Random Forest (RF), Linear Discriminant Analysis (LDA), k-nearest neighbor (KNN), Adaboost (AB), Support Vector Machine (SVM), and XG Boost (XGB). All FB models were created using scikit-learn 1.2.1 ML library.

To train and evaluate each FB model, models were first initialized for each of the chosen algorithms, calculating performance metrics and feature importances directly during model evaluation. The training process involved setting up specific models, such as SVM, using the scikit-learn's fit method. A 5-fold cross-validation method was utilized. The dataset was randomly divided into equal five test sets. For each of the test sets, the rest of the data were split into training (80%) and validation (20%) sets. For each of the architectures, 5-fold cross-validation and 5 training iterations were performed. Comparisons of these predictions with the actual labels facilitated the calculation of key performance metrics including accuracy and F1 score, averaged across the 5 folds.

Further, we generated a feature importance list for each model, which ranked the features by their influence on the model's predictions. This provided critical insights into which variables significantly impacted the detection of anxiety. Detailed results for each model's performance, along with the top seven influential features, were systematically documented to underscore the predictive capabilities and strengths of each model within the study.

2.6. Transfer Learning

Afterwards, models that were trained using a subset of the WESAD dataset with and without noise augmentation that matched in modalities and were similar in locations were tested on the three subsets to test the validity of transfer learning for this application. Feature-based models were employed. DT, RF, LDA, KNN, AB, SVM, and XGB were utilized to test performance for transfer learning, using re-used models without additional fine-tuning. As seen in Figure 3, ML noise-augmented data were used to train feature-based models that were directly applied to the three subsets of RADWear and WEAR data.

Noise-augmented data derived from the WESAD dataset was used to assess the resilience of machine learning models to environmental disturbances. This approach was crucial for understanding how these models performed under simulated conditions that mimicked real-world noise, which is often encountered in daily activities and can significantly affect the accuracy of anxiety detection systems. Noise augmentation was carried out by adding Gaussian noise, ranging from 0.0001 to 0.6 signal-to-noise ratios (SNRs) from WESAD data.

By training models on both the original and noise-augmented WESAD data, we established a performance benchmark, examining how noise impacts model effectiveness and identifying which models maintain their predictive power despite increased noise levels. This process not only highlighted the robustness of certain models but also allowed for the optimization of these models to withstand typical real-world disruptions.

The transfer learning techniques applied here involved testing the models—initially trained on WESAD and its noise-augmented version—on the RADWear and WEAR data. The transition to testing on these studies, which feature real-life conditions and more complex environments than those simulated by noise augmentation, was facilitated by the preliminary insights gained from the noise impact analysis. This step ensured that the models not only generalized well across different types of input data but were also applicable in practical settings where environmental variability is the norm.

Overall, analyzing models with noise-augmented data enhances the relevance and applicability of these models for RADWear and WEAR data, where real-world conditions play a crucial role. It provided a solid foundation for understanding model performance in the face of unpredictable environmental factors, ensuring that the anxiety detection systems developed are both effective and reliable in varied real-world scenarios.

For transfer learning, models that have been tested on WESAD, and noise-augmented WESAD, were tested on the three subsets. The use of WESAD serves to establish a performance benchmark as a reference point. Since the models were initially trained and validated on this dataset, showcasing their performance on WESAD helps in understanding their efficiency and accuracy before they were tested under more variable conditions such as those provided by the RADWear and WEAR datasets. It is important to articulate that the WESAD dataset acts as a foundational dataset for training the models, whose performance metrics were crucial for establishing a comparative analysis.

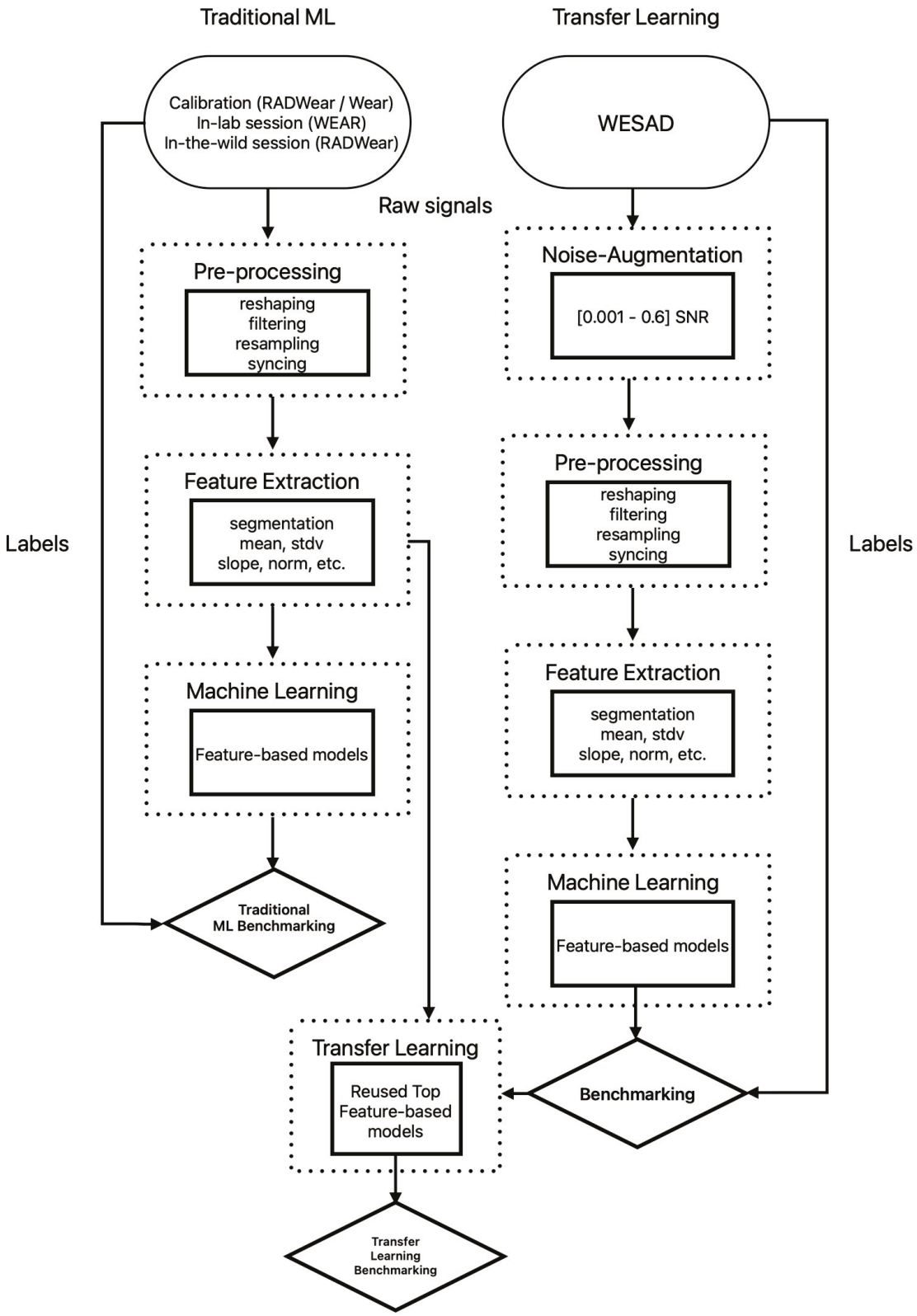


Figure 3. Flowchart outlining pipeline for processing and analyzing wearable sensor data from raw data collection to machine learning.

The incorporation of noise-augmented data was a strategic decision to bolster the resilience of machine learning models against environmental noise, which is a common and disruptive element in real-world settings. The process of augmenting the WESAD dataset with noise simulates these realistic disturbances, thus enabling us to stress-test

the models and prepare them for the unpredictable variances they would encounter in practical scenarios. By doing so, we aimed to ensure that the models retain their predictive accuracy even when faced with data that has been compromised or distorted by external factors. To assess the performance of each model evaluated at each subset, F_1 score and accuracy were used to evaluate the models.

3. Results

3.1. Traditional ML Models

The traditional ML models refer to those tested and trained exclusively on each of the three data subsets, including the original WESAD dataset as a reference point for benchmarking. The analysis of the model performances with the RADWear and WEAR data (Table 3) highlighted the excellent performance of the XGB and DT models. Notably, the XGB model achieved a 0.99 accuracy and F1 score in the WESAD dataset [22]. This performance was maintained across the RADWear and WEAR calibration subsets, where XGB recorded a 0.95 accuracy and a 0.94 F1 score. Furthermore, even in the challenging conditions of the WEAR in-lab settings, XGB showed a robust performance with a 0.92 accuracy and a 0.90 F1 score. These results validate the effectiveness of traditional ML models in real-world scenarios, substantially outperforming the 50% accuracy expected of random guessing in binary classification.

Table 3. F1-score performance results of traditional ML models trained and tested on the same subset of the datasets.

Data	XGB		DT		LDA		RF	
	ACC	**F_1**	**ACC**	**F_1**	**ACC**	**F_1**	**ACC**	**F_1**
WESAD	0.99	0.99	0.99	0.99	0.93	0.92	0.91	0.89
RADWear + WEAR Calibration	0.95	0.94	0.91	0.9	0.82	0.8	0.88	0.86
WEAR in-lab	0.92	0.9	0.93	0.94	0.69	0.65	0.8	0.76
RADWear in-the-wild (balanced to 30%)	0.87	0.77	0.97	0.95	0.79	0.52	0.82	0.72
RADWear in-the-wild (balanced to 41%)	0.88	0.58	0.99	0.66	0.73	0.61	0.78	70
Random guess	0.5							

Note: Grey color used to indicate top performing classifier, based on accuracy (ACC) or F1 score (F_1), for each dataset.

Interestingly, the most predictive features varied across models and datasets. In Table 4, the performance metrics of various features across different models are presented, offering a detailed comparison of how different data types influence the effectiveness of the models. The "Weighted average" represents the aggregated influence of each feature across different models. This calculation was performed using a weighted mean, where each feature's importance is squared within a model, summed across models, and then divided by the total sum of the feature importances. This calculation method emphasizes features that were consistently significant across various models, highlighting their predictive power in detecting anxiety. This approach particularly underscores the adaptability and critical role of specific features under conditions such as motion, where accelerometer data becomes increasingly significant.

Table 4. Feature importance for traditional ML models on calibration, in-lab, and in-the-wild data.

	Modality	Modality Weighted	Feature	Feature Weighted	RF	DT	XGB
RADWear in-the-wild	ECG	0.13	ECG_{min}	0.17	0.14	0.18	0.19
			ECG_{std}	0.08	0.08		
			ECG_{max}	0.06	0.06	0.04	0.07
	Resp	0.07	$Resp_{Exhal\ std}$	0.07	0.07		
			$Resp_{mean}$	0.07	0.07	0.09	0.04
			$Resp_{rate}$	0.06	0.07		0.06
			$Resp_{I/E}$	0.06	0.06		
			$Resp_{min}$	0.08		0.08	0.08
	EDA	0.06	EDA_{mean}	0.07		0.07	
			EDA_{drange}	0.06			0.06
	BVP	0.04	$BVP_{peak\ freq}$	0.04			0.04
	TEMP	0.11	$TEMP_{mean}$	0.11		0.11	
	ACC	0.06	$ACC_{net\ w\ mean}$	0.06		0.06	
RADWear + WEAR Calibration	EDA	0.1	EDA_{mean}	0.12	0.06	0.17	0.05
			EDA_{max}	0.05			0.05
	RESP	0.05	$Resp_{min}$	0.07		0.07	
			$Resp_{mean}$	0.05	0.05		0.05
			$Resp_{rate}$	0.05		0.05	0.04
			$Resp_{Inhalmean}$	0.04		0.03	0.04
	BVP	0.04	BVP_{max}	0.04	0.04		
	ACC	0.13	$ACC_{net\ w\ std}$	0.21	0.08	0.26	0.2
			$ACC_{net\ w\ mean}$	0.04		0.04	
			$ACC_{x\ min}$	0.06		0.06	
			$ACC_{x\ C\ std}$	0.08	0.08		
			$ACC_{x\ min}$	0.06	0.07		0.04
			$ACC_{x\ mean}$	0.05	0.05		
WEAR in-lab	ECG	0.12	ECG_{bpm}	0.08			0.08
			ECG_{max}	0.13	0.14	0.14	0.12
			ECG_{min}	0.07	0.07		
			ECG_{std}	0.15	0.2	0.12	0.07
	EDA	0.05	EDA_{max}	0.05		0.06	0.04
			EDA_{mean}	0.05		0.05	
	RESP	0.06	$Resp_{mean}$	0.07	0.08	0.09	0.05
			$Resp_{min}$	0.04	0.04		0.05
	ACC	0.05	$ACC_{x\ mean}$	0.05		0.06	0.05
			$ACC_{net\ w\ mean}$	0.05	0.05		
	TEMP	0.14	$TEMP_{mean}$	0.14	0.12	0.16	

Note: Grey color used to indicate most important modality and feature for each dataset and classifier.

3.2. Feature-Based Transfer Learning

When evaluating the transfer learning performance (Figure 4 and Table 5), we discovered that the RF model excelled in the calibration and in-lab settings, while SVM achieved the best results for in-the-wild data. Intriguingly, models like XGB and DT, which previously outperformed in the baseline case, did not maintain their superiority in the transfer learning scenario. This finding highlights the importance of carefully selecting models based on the specific characteristics of the target dataset and the potential limitations of directly transferring models trained on one dataset to another.

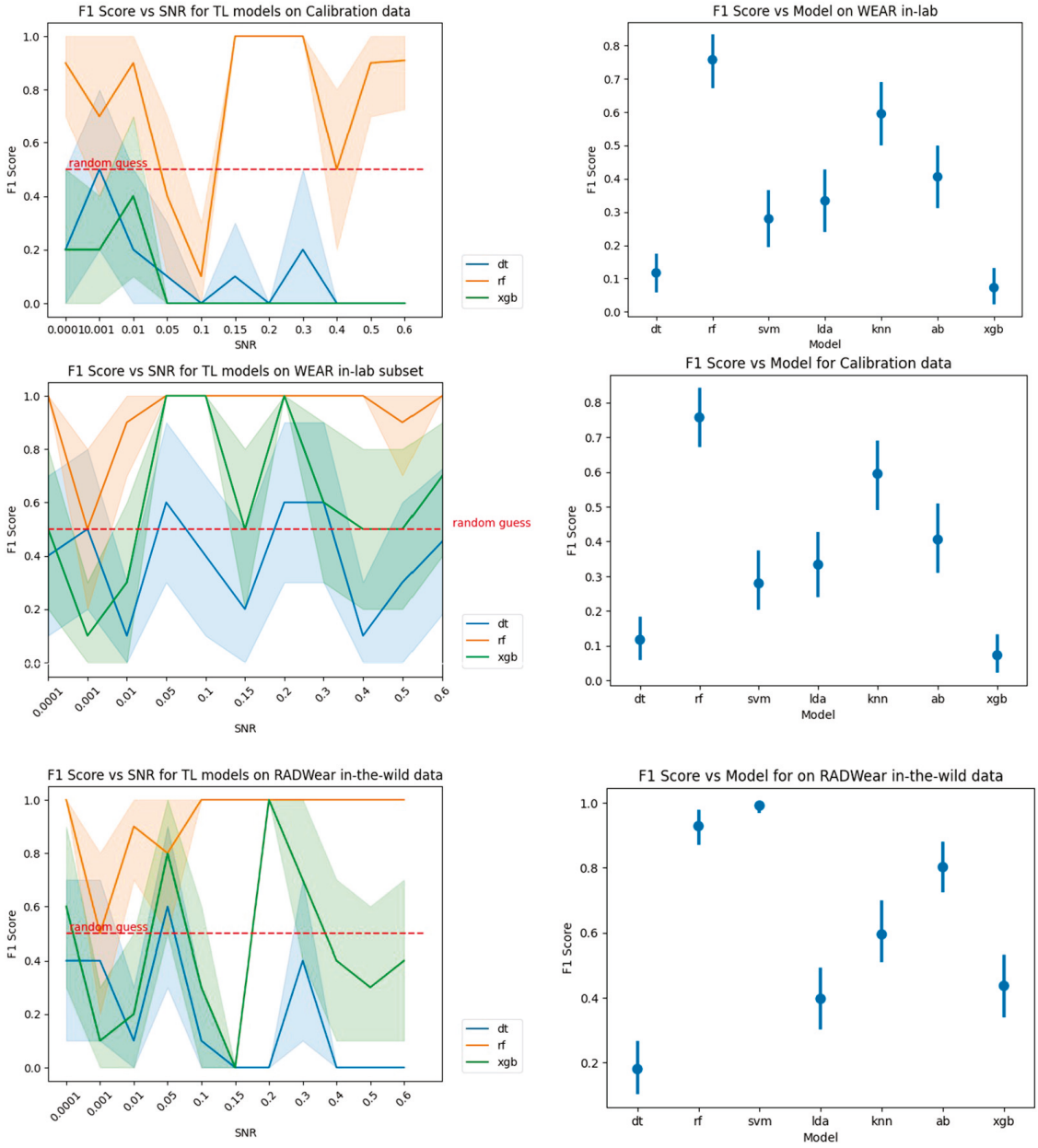


Figure 4. The F1 score of transfer learning models are shown with the left column showing F_1 score for models trained at different noise levels at different signal-to-noise ratios (SNRs), shown in the x-axis. The right column shows F_1 score for each of the tested models for all SNRs. F_1 score, as tested on calibration data for WEAR and RADWear in the top row, WEAR in-lab sessions in the middle row, and on RADWear in-the-wild in the bottom row.

Table 5. Performance results for transfer learning models on calibration, in-lab, and in-the-wild data.

Model	DT		RF		XGB	
	Accuracy	**F_1-Score**	**Accuracy**	**F_1-Score**	**Accuracy**	**F_1-Score**
RADWear + WEAR Calibration	0.12	0.12	0.76	0.76	0.07	0.07
WEAR in-lab	0.18	0.39	0.93	0.94	0.44	0.61
RADWear in-the-wild anxiety class balanced to 41%	0.18	0.18	0.93	0.93	0.44	0.44

In exploring the performance of transfer learning across our datasets, we identified a significant trend related to the distribution of feature importance weights in the models. Models exhibiting more evenly distributed feature importance weights demonstrated more successful adaptation when applied to new datasets, particularly when transitioning from controlled to more variable real-world conditions. For instance, models like LDA, which generally maintained balanced importance across features, adapted more effectively compared to those heavily reliant on one or two specific features, such as XGB. The latter showed reduced performance during transfer learning tasks, especially when applied to the in-the-wild subset of the WEAR dataset. This suggests that a more uniform distribution of feature weights might enhance a model's adaptability, supporting better generalization across different experimental conditions and datasets. This finding underscores the importance of considering feature balance during model training for effective transfer learning applications.

4. Discussion

Our study contributes to the field of anxiety detection using wearables and machine learning by demonstrating the feasibility of developing robust models that can accurately detect anxiety in noisy environments and can be transferred from controlled laboratory conditions. The strong performance of our models, particularly XGB and DT, in both controlled and real-world settings, consistent with recent work in open-source datasets [22], underscores the potential for deploying such techniques in practical mental health monitoring applications.

In comparison to prior related work [14,19] examining physiological signals (Table 6), we find that the XGB and DT classifiers provide improved accuracy across all datasets examined in this study when using test data from the same dataset. Similarly, the F1 scores for the DT classifier demonstrated improved performance across all datasets except the in-the-wild data balanced to 41%. When examining transfer learning, by evaluating the use of training on one dataset and comparing to another, we found the best performance was provided by an RF classifier, consistent with prior work, which found a RF anxiety classifier to demonstrate robust performance when applying to a new dataset [19].

Table 6. Anxiety classification performance results of related works using physiological signals.

Authors	Dataset Used	Features	Psychological Measures	Physical Activity or Condition	Environmental Noise	Classification/Cut-Off	Performance (Accuracy)	Performance (F1)	TL Performance (Accuracy)	TL Performance (F1)
Henry et al., 2023 [19]	CASE: customized anxiety index from self-report arousal and valance scores	ECG and BVP	Anxiety index	Seated	Fixed	Anxiety State, using anxiety index with 0.5 cutoff	RF(BVP) = 0.752; SVM(ECG) = 0.763.	RF(BVP) = 0.759; SVM(ECG) = 0.770.	WESAD to CASE: SVM, RF, XGBoost, MLP, ResNet(BVP) = 0.501; XGBoost(ECG) = 0.605.	WESAD to CASE:FTTP(BVP) = 0.107; FTTP(ECG) = 0.554.
	WESAD: Study protocol design of baseline, amusement, stress	ECG and BVP	PANAS, STAI, SAM, and SSSQ	Seated/Standing	Fixed	Stress vs. Non-Stress using protocol label	SVM(BVP) = 0.745; SVM(ECG) = 0.811.	SVM(BVP) = 0.773; SVM(ECG) = 0.818.	CASE to WESAD: MLP(BVP) = 0.587; RF(ECG) = 0.657.	CASE to WESAD: SVM(BVP) = 0.694; FTT(ECG) = 0.694.
Perpetuini et al., 2020 [14]	Study protocol examined supine rest with and without video	BVP	STAI	Supine	Fixed	Anxiety State, using STAI with 40 cutoff	AUC GLM(BVP) = 0.88;	NA	NA	NA

Prior work using the WESAD dataset has focused on stress detection using ML and deep learning approaches [8,34–36]. In addition, the in-lab portions of the WEAR study have been primarily focused on establishing the validity of physiological signal changes in stressful conditions [31]. However, further work remains to validate the in-the-wild portions of the RADWEAR study.

This study advances the existing literature in several ways. First, we conducted an analysis of feature importance, shedding light on the relative contributions of different physiological signals and derived features in anxiety detection, which could be potentially beneficial for individualized treatment planning. Second, we explored the applicability of transfer learning, highlighting the challenges and opportunities in adapting models trained on one dataset to another. Notably, it seems that models with feature importances close in weight perform better for transfer learning, while models with feature importances heavily weighted on one or two features, like XGB for calibration and in-the-wild, do not perform transfer learning as well.

The contrast in the identification of the most important modalities across each of the datasets suggests that physical activity and environmental noise significantly impact the most important information to classify anxiety. In contrast to prior work examining data from controlled environments [19,22], where EDA and BVP have been found to be among the most important modalities for anxiety classification, we found that ACC, TEMP, and ECG provided the most important modalities overall, as physical activity and environmental noise increased.

The study's findings emphasize the importance of selecting features that were robust to various environmental factors and failure modes. Consequently, models that rely on a distributed feature importance, where multiple signals contribute to the detection of anxiety, were more likely to be resilient to individual signal failures. Future research should focus on identifying and engineering features that are robust to these challenges, enhancing the reliability of anxiety detection models in real-world settings.

However, our study is not without limitations. The sample sizes of the RADWear and WEAR datasets used in this study, while diverse, may not fully capture the entire spectrum of individual variability in anxiety responses. As those studies are ongoing, these datasets will expand, incorporating a broader range of participant demographics and clinical profiles, so as to better generalize new subjects. Additionally, while our feature-based models demonstrate promising performance, further validation in longitudinal studies and real-world deployments would be necessary to assess their long-term reliability and usefulness in clinical practice. Further, the high accuracies but lower F1 scores observed in some of the transfer learning suggest potential overfitting, which should be further examined in future work.

Developing accurate and robust anxiety detection models can pave the way for personalized mental health interventions that are delivered in real-time using wearable devices, potentially enhancing the effectiveness of current empirically supported therapy, improving the accuracy of self-report data, and providing "self-help" programs beyond common bibliotherapy, thereby reaching more rural populations where specialized clinicians may not be available. This could revolutionize the way we monitor and manage anxiety disorders, enabling early detection, timely support, and targeted treatments. However, translating these models into practical tools will require addressing challenges such as data privacy, user acceptance, and seamless integration with existing healthcare systems.

5. Conclusions

This study underscores the efficacy of feature-based models, particularly XGBoost and DT, in accurately detecting anxiety under both controlled conditions and real-world

environments, and for models such as RF to transfer learning well from different datasets. The introduction of the RADWear and WEAR studies enhances our understanding of the robustness of these models amidst environmental noise and varied conditions. The analysis of feature importance and the implementation of transfer learning have significantly contributed to advancing the field of anxiety detection using wearable technology. Key findings emphasize the resilience of feature-based models and the critical role of precise feature selection in maintaining model accuracy across diverse settings. These insights not only validate the effectiveness of current methodologies but also underscore the potential of these models in practical mental health monitoring applications.

Author Contributions: Conceptualization, A.A., J.C. (Jean Clore), R.S., E.T.H.-W., and M.E.H.; methodology, A.A., J.C. (Jonathan Cerna), and M.H.; software, A.A.; validation, A.A.; formal analysis, A.A.; resources, A.A.; data curation, A.A., M.H., and J.C. (Jonathan Cerna); writing—original draft preparation, A.A., J.C. (Jean Clore and Jonathan Cerna), R.S., E.T.H.-W., and M.E.H.; visualization, A.A.; supervision, E.T.H.-W. and M.E.H.; project administration, M.E.H.; funding acquisition, J.C. (Jean Clore), R.S., E.T.H.-W., and M.E.H. All authors have contributed to writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Jump ARCHES endowment through the Health Care Engineering Systems Center and partly supported by the ACCESS MATCH program made possible by the U.S. National Science Foundation.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board (or Ethics Committee) of the University of Illinois Urbana-Champaign (protocol 21769, approved on 1 December 2022) and University of Illinois College of Medicine at Peoria (protocol 1880297, approved on 27 April 2022).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Acknowledgments: We thank members of the Mobility and Fall Prevention Research Lab and the Human Dynamics and Control Lab that contributed to this study, and participants for making this study possible.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Celano, C.M.; Daunis, D.J.; Lokko, H.N.; Campbell, K.A.; Huffman, J.C. Anxiety Disorders and Cardiovascular Disease. *Curr. Psychiatry Rep.* **2016**, *18*, 101. [CrossRef] [PubMed]
2. Critchley, H.D. Study of the Stress Response: Role of Anxiety, Cortisol and DHEAs. *Neuroscientist* **2002**, *8*, 132–142. [CrossRef]
3. Bär, K.J.; Critchley, H. Autonomic Control. *Brain Mapp. Encycl. Ref.* **2015**, *2*, 635–642. [CrossRef]
4. Julian, L.J. Measures of Anxiety. *Arthritis Care* **2011**, *63*, 20561. [CrossRef]
5. Healey, J.A.; Picard, R.W. Detecting Stress during Real-World Driving Tasks Using Physiological Sensors. *IEEE Trans. Intell. Transp. Syst.* **2005**, *6*, 156–166. [CrossRef]
6. Elgendi, M.; Galli, V.; Ahmadizadeh, C.; Menon, C. Dataset of Psychological Scales and Physiological Signals Collected for Anxiety Assessment Using a Portable Device. *Data* **2022**, *7*, 132. [CrossRef]
7. Haouij, N.E.; Poggi, J.M.; Sevestre-Ghalila, S.; Ghozi, R.; Jadane, M. Affective ROAD System and Database to Assess Driver's Attention. In Proceedings of the 33rd Annual ACM Symposium on Applied Computing, Pau, France, 9–13 April 2018; Association for Computing Machinery: New York, NY, USA, 2018; pp. 800–803. [CrossRef]
8. Schmidt, P.; Reiss, A.; Duerichen, R.; Laerhoven, K. Van Introducing WeSAD, a Multimodal Dataset for Wearable Stress and Affect Detection. In Proceedings of the 2018 International Conference on Multimodal Interaction, Boulder, CO, USA, 16–20 October 2018; Association for Computing Machinery: New York, NY, USA, 2018; pp. 400–408. [CrossRef]

9. Shaukat-Jali, R.; van Zalk, N.; Boyle, D.E. Detecting Subclinical Social Anxiety Using Physiological Data from a Wrist-Worn Wearable: Small-Scale Feasibility Study. *JMIR Form. Res.* **2021**, *5*, 32656. [CrossRef]
10. Sanjay, V.M.; Ankith, I.; Dhanush, G.C.; Akshaya, H.P.; Malik, J.; Tejaswini, B.N. Anxiety Prediction during Stressful Scenarios Using Machine Learning. In Proceedings of the 2022 6th International Conference on Intelligent Computing and Control Systems, ICICCS 2022, Madurai, India, 25–27 May 2022; pp. 1199–1202.
11. Albagmi, F.M.; Alansari, A.; Shawan, D.S.A.; AlNujaidi, H.Y.; Olatunji, S.O. Prediction of Generalized Anxiety Levels during the Covid-19 Pandemic: A Machine Learning-Based Modeling Approach. *Inf. Inform. Med. Unlocked* **2022**, *28*, 100854. [CrossRef]
12. Ramani, B.; Patel, W.; Solanki, K. Stress Ocare: An Advance IoMT Based Physiological Data Analysis for Anxiety Status Prediction Using Cloud Computing. *J. Discret. Math. Sci. Cryptogr.* **2022**, *25*, 1019–1029. [CrossRef]
13. Srinath, K.S.; Kiran, K.; Pranavi, S.; Amrutha, M.; Shenoy, P.D.; Venugopal, K.R. Prediction of Depression, Anxiety and Stress Levels Using Dass-42. In Proceedings of the 2022 IEEE 7th International conference for Convergence in Technology, I2CT 2022, Pune, India, 7–9 April 2022; pp. 1–6.
14. Perpetuini, D.; Chiarelli, A.M.; Cardone, D.; Filippini, C.; Rinella, S.; Massimino, S.; Bianco, F.; Bucciarelli, V.; Vinciguerra, V.; Fallica, P.; et al. Prediction of State Anxiety by Machine Learning Applied to Photoplethysmography Data. *PeerJ* **2021**, *9*, e10448. [CrossRef]
15. Ancillon, L.; Elgendi, M.; Menon, C. Machine Learning for Anxiety Detection Using Biosignals: A Review. *Diagnostics* **2022**, *12*, 1794. [CrossRef] [PubMed]
16. Sato, M.; Ishikawa, W.; Suzuki, T.; Matsumoto, T.; Tsujii, T.; Sakatani, K. Bayesian STAI Anxiety Index Predictions Based on Prefrontal Cortex NIRS Data for the Resting State. *Adv. Exp. Med. Biol.* **2013**, *765*, 251–256. [CrossRef] [PubMed]
17. Baird, A.; Cummins, N.; Schnieder, S.; Krajewski, J.; Schuller, B.W. An Evaluation of the Effect of Anxiety on Speech ? Computational Prediction of Anxiety from Sustained Vowels. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, Shanghai, China, 25–29 October 2020; Volume 2020, pp. 4951–4955.
18. Fukuda, Y.; Ishikawa, W.; Kanayama, R.; Matsumoto, T.; Takemura, N.; Sakatani, K. Bayesian Prediction of Anxiety Level in Aged People at Rest Using 2-Channel NIRS Data from Prefrontal Cortex. *Adv. Exp. Med. Biol.* **2014**, *812*, 303–308. [CrossRef] [PubMed]
19. Henry, J.; Lloyd, H.; Turner, M.; Kendrick, C. On the Robustness of Machine Learning Models for Stress and Anxiety Recognition from Heart Activity Signals. *IEEE Sens. J.* **2023**, *23*, 14428–14436. [CrossRef]
20. Xiong, H.; Berkovsky, S.; Romano, M.; Sharan, R.V.; Liu, S.; Coiera, E.; McLellan, L.F. Prediction of Anxiety Disorders Using a Feature Ensemble Based Bayesian Neural Network. *J. Biomed. Inf. Inform.* **2021**, *123*, 103921. [CrossRef]
21. Adheena, M.A.; Sindhu, N.; Jerritta, S. Physiological Detection of Anxiety. In Proceedings of the 2018 International Conference on Circuits and Systems in Digital Enterprise Technology (ICCSDET), Kottayam, India, 21–22 December 2018; IEEE: Piscataway Township, NJ, USA, 2018; pp. 1–5.
22. Alkurdi, A.; Clore, J.; Sowers, R.; Hsiao-Wecksler, E.T.; Hernandez, M.E. Resilience of Machine Learning Models in Anxiety Detection: Assessing the Impact of Gaussian Noise on Wearable Sensors. *Appl. Sci.* **2025**, *15*, 88. [CrossRef]
23. Pan, S.J.; Yang, Q. A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng.* **2010**, *22*, 1345–1359. [CrossRef]
24. Singh, M.; Prakash, P.; Kaur, R.; Sowers, R.; Brašić, J.R.; Hernandez, M.E. A Deep Learning Approach for Automatic and Objective Grading of the Motor Impairment Severity in Parkinson's Disease for Use in Tele-Assessments. *Sensors* **2023**, *23*, 9004. [CrossRef]
25. Weiss, K.; Khoshgoftaar, T.M.; Wang, D.D. A Survey of Transfer Learning. *J. Big Data* **2016**, *3*, 9. [CrossRef]
26. Zhuang, F.; Qi, Z.; Duan, K.; Xi, D.; Zhu, Y.; Zhu, H.; Xiong, H.; He, Q. A Comprehensive Survey on Transfer Learning. *Proc. IEEE* **2021**, *109*, 43–76. [CrossRef]
27. Shao, L.; Zhu, F.; Li, X. Transfer Learning for Visual Categorization: A Survey. *IEEE Trans. Neural Netw. Learn. Syst.* **2015**, *26*, 1019–1034. [CrossRef]
28. Watson, D.; Clark, L.A.; Tellegen, A. Development and Validation of Brief Measures of Positive and Negative Affect: The PANAS Scales. *J. Pers. Soc. Psychol.* **1988**, *54*, 1063–1070. [CrossRef]
29. Spielberger, C.D.; Gonzalez-Reigosa, F.; Martinez-Urrutia, A.; Natalicio, L.F.S.; Natalicio, D.S. The State-Trait Anxiety Inventory. *Interam. J. Psychol.* **1971**, *5*, 3–4.
30. Helton, W.S. Validation of a Short Stress State Questionnaire. *Proc. Hum. Human. Factors Ergon. Soc. Annu. Meet.* **2004**, *48*, 1238–1242. [CrossRef]
31. He, M.; Cerna, J.; Alkurdi, A.; Dogan, A.; Zhao, J.; Clore, J.L.; Sowers, R.; Hsiao-Wecksler, E.T.; Hernandez, M.E. Physical, Social and Cognitive Stressor Identification Using Electrocardiography-Derived Features and Machine Learning from a Wearable Device. In Proceedings of the 2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Orlando, FL, USA, 15–19 July 2024; pp. 1–4.
32. Malik, M.; Camm, A.J.; Bigger, J.T.; Breithardt, G.; Cerutti, S.; Cohen, R.J.; Coumel, P.; Fallen, E.L.; Kennedy, H.L.; Kleiger, R.E.; et al. Heart Rate Variability: Standards of Measurement, Physiological Interpretation, and Clinical Use. *Circulation* **1996**, *93*, 1043–1065. [CrossRef]

33. Scholkmann, F.; Boss, J.; Wolf, M. An Efficient Algorithm for Automatic Peak Detection in Noisy Periodic and Quasi-Periodic Signals. *Algorithms* **2012**, *5*, 588–603. [CrossRef]
34. Gil-Martin, M.; San-Segundo, R.; Mateos, A.; Ferreiros-Lopez, J. Human Stress Detection with Wearable Sensors Using Convolutional Neural Networks. *IEEE Aerosp. Electron. Syst. Mag.* **2022**, *37*, 60–70. [CrossRef]
35. Ghosh, S.; Kim, S.; Ijaz, M.F.; Singh, P.K.; Mahmud, M. Classification of Mental Stress from Wearable Physiological Sensors Using Image-Encoding-Based Deep Neural Network. *Biosensors* **2022**, *12*, 1153. [CrossRef]
36. Dziezyc, M.; Gjoreski, M.; Kazienko, P.; Saganowski, S.; Gams, M. Can We Ditch Feature Engineering? End-to-End Deep Learning for Affect Recognition from Physiological Sensor Data. *Sensors* **2020**, *20*, 6535. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

MDPI AG
Grosspeteranlage 5
4052 Basel
Switzerland
Tel.: +41 61 683 77 34

Sensors Editorial Office
E-mail: sensors@mdpi.com
www.mdpi.com/journal/sensors



Disclaimer/Publisher's Note: The title and front matter of this reprint are at the discretion of the Guest Editors. The publisher is not responsible for their content or any associated concerns. The statements, opinions and data contained in all individual articles are solely those of the individual Editors and contributors and not of MDPI. MDPI disclaims responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.





Academic Open
Access Publishing

mdpi.com

ISBN 978-3-7258-8548-0