

Article

From Signal to Semantics: The Multimodal Haptic Informatics Index for Triangulating Haptic Intent at the Edge

Song Xu ^{1,2} , Chen Li ² , Jia-Rong Li ¹  and Teng-Wen Chang ^{1,*} 

¹ Master & Doctoral Program, Graduate School of Design, National Yunlin University of Science and Technology, Yunlin 64002, Taiwan; d10430023@yuntech.edu.tw (S.X.); d11430002@yuntech.edu.tw (J.-R.L.)

² School of Design and Art, Beijing Institute of Technology, Zhuhai 519088, China; lichen@bitzh.edu.cn

* Correspondence: tengwen@softlab.tw or tengwen@yuntech.edu.tw; Tel.: +886-5-534-2601 (ext. 6510)

Abstract

Modern interaction with smart devices is hindered by the “Midas Touch” problem, where sensors frequently misinterpret incidental physical movements as intentional commands due to a lack of human context. This research addresses this conflict by introducing the Multimodal Haptic Informatics (MHI) index within a novel Scene–Action–Trigger (SAT) framework. The goal is to contextualize mechanical movements as human intent by integrating physical, spatial, and cognitive data locally at the edge. The methodology employs an “Action-as-primary indexing” mechanism where the Action channel (IMU) serves as a temporal anchor t , triggering high-resolution Scene (computer vision) and Trigger (audio) processing only during critical haptic events. Validated through a complex origami crane task generating 29,408 data frames, the framework utilizes a three-stage informatics derivation process: single-modal scoring, score weighting, and hand state mapping. Results demonstrate that applying an adaptive “Speedometer” logic successfully reclassifies the “Transitional State”. While this state constitutes over half of the behavioral dataset (54.76% on average), it is effectively disambiguated into meaningful intent using a self-trained local Large Language Model (LLM) for semantic verification. Furthermore, the event-driven sampling of 93 keyframes reduces the processing overhead by 99.68% compared to linear annotation. This study contributes a low-latency, privacy-preserving “Protocol of Assent” that maintains user agency by providing intelligent system suggestions based on confirmed haptic intensity.

Keywords: Midas Touch; wearable sensing; tiny machine learning; context awareness; design informatics



Academic Editor: Stefanos Kollias

Received: 20 January 2026

Revised: 9 February 2026

Accepted: 13 February 2026

Published: 15 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

1.1. The Challenge of Intent Recognition in Modern HCI

The contemporary landscape of human–computer interaction (HCI) is undergoing a significant transition from explicit, command-driven inputs toward continuous and pervasive sensing environments. As interactive systems become increasingly integrated into the fabric of daily life, there is a growing expectation for these devices to exhibit proactive and context-aware behaviors. This paradigm shift necessitates a move beyond traditional kinematic classification, which merely identifies physical movement, toward intent-aware sensing that can interpret the underlying purpose of human actions [1]. This challenge is particularly acute in specialized manual tasks, such as paper folding or architectural modeling, where physical behaviors are fluid, overlapping, and highly

variable [2]. This complexity necessitates a paradigm shift in how design actions are coded and quantified. Traditional linkography techniques often struggle to scale in these fluid environments due to the time-consuming nature of manual annotation [3]. To address this, “Fuzzy Linkography” provides a means of automatically producing creative activity traces through computational models of semantic similarity, enabling the classification of user behaviors at scale [4]. Furthermore, the significance of such automated representations lies in their ability to support qualitative analysis of design framing in near real-time, bridging the gap between design science and practice [5]. Within such environments, the mapping of user experiences becomes essential for creating effective interfaces as demonstrated by the “Sensible Energy System” (SENS), which facilitates interaction between behavior and resource consumption in co-making spaces. Such frameworks provide dynamic feedback on energy consumption patterns by predicting energy outputs and optimizing management through spatial interaction analysis [6].

Moreover, modern design thinking increasingly relies on a “seeing–moving–seeing” framework, enabling designers to convey intuitive ideas through direct observation and representations in design media. This process effectively connects physical manufacture with digital interaction, reducing the disparity between virtual and physical fabrication while ensuring that computing systems remain aligned with human behavior [7]. Furthermore, the exploration of semantic–visual stimuli through adaptive workflows supports continuous ideation by extracting and summarizing features from design templates. This engineering of design knowledge provides necessary hints for emerging pathways of insight, bridging the gap between informatics and creative problem-solving [8].

Despite the high capabilities of 6-axis Inertial Measurement Units (IMUs) in providing fine-grained gesture recognition on wearable platforms [9], these sensors often operate within a “semantic vacuum”. This state occurs when high-fidelity kinematic data lacks the human context required to confirm a legitimate cognitive intent [10]. The persistent “ambiguity dilemma” further complicates the robust recognition of intent, as identical kinematic signals can often map to multiple postures or different stages of a process, leading to significant pose ambiguity [11,12]. Even when performing the same gesture, variations in amplitude or the identity of the user can result in recognition failure [13]. To mitigate these issues, digital methods can be employed to perform multi-scale analysis of the design process, treating viewing behavior as structured cognitive networks [14,15]. The integration of situated Function–Behavior–Structure (FBS) models with linkography further allows researchers to unravel complex cognitive patterns during co-design, ensuring that digital systems remain user-centric [16]. While these cognitive frameworks provide a theoretical lens for understanding co-design, the current haptic sensing hardware remains largely disconnected from such high-level structures. This lack of a semantic anchor in kinematic-only streams makes it difficult to distinguish between purposeful interactions and incidental movement [17]. Although prior attempts at mitigation have utilized sensor fusion, these methods frequently reach their limits in real-world contexts where user behavior is highly stochastic [18].

Responsive structures must adapt to external stimuli by incorporating dynamic forms, yet the learning curve for operating such sophisticated design tools remains a primary hurdle for modern designers [14]. Consequently, this study proposes a novel edge-computing architecture for multimodal intent recognition, introducing a quantifiable index (MHI) for haptic intent characterization. This research establishes a “Protocol of Assent” where digital systems require multimodal consistency before confirming a haptic intent, providing a privacy-preserving solution to the Midas Touch problem. By automating the interaction analysis between sketching and digital fabrication environments, these models help identify the cognitive gaps that impact productivity in smart design ecosystems [19]. While the

landscape of intent recognition is vast, this research focuses its initial inquiry on fine-motor haptic intent within creative design settings, intentionally delimiting the scope from gross motor skill assessments and routine daily life tasks to ensure a high-fidelity baseline for multimodal triangulation.

1.2. Midas Touch Issue: Misalignment Between Signals and Intentions

The “Midas Touch” (MT) problem represents a significant obstacle in the design of non-conventional user interfaces, particularly those relying on pointers and gestural selection components [20]. It is defined as the occurrence of involuntary selection actions triggered by natural human movement during system interaction [21]. In scenarios where a designer is engaged in a fluid task like paper folding, the hands are in constant motion for exploration and adjustment [2]. Traditional algorithms often fail to disambiguate these exploratory behaviors from purposeful selection commands, leading to incidental motions being misinterpreted as active digital actions. This misalignment between physical signals and cognitive intentions significantly reduces user agency and system reliability, as users are forced to move with excessive caution to avoid accidental triggers [20].

As sensor sensitivity increases to capture more fine-grained movements, the frequency of such accidental triggers typically rises, further degrading the user experience. In gesture-based systems, the absence of a clear physical “click” means the system must rely on imperfect kinematic patterns to signify intent [22]. To mitigate this, specific interaction methods like head-tracking bars or the “Pactolo Bar” have been developed to enable or disable click events on demand [20]. Furthermore, the presence of sensor position drift in wearable devices can shift the system’s baseline for what constitutes a “command”, increasing the likelihood of misinterpretation [23].

To solve these selection ambiguities, researchers have explored various gesture sequences, such as two-step taps or double-taps, to ensure intentionality [21]. Moreover, user-independent gesture recognition models utilizing sensor fusion—specifically combining Electromyography (EMG) and IMU data—have been studied as a means to create more natural interfaces. By fusing muscle activity with motion data, systems can achieve higher classification accuracies and reduce calibration times for new users [24]. Beyond technical malfunctions, the widespread adoption of wearable sensing technology may also pose a cognitive threat. Users may feel uncomfortable wearing conspicuous or overly sensitive technology. This psychological factor is often linked to a lack of aesthetic appeal or the neglect of privacy, leading users to abandon technology out of frustration [25]. Addressing these design factors is essential for ensuring that interactive devices are accepted and respect user autonomy. In addressing the selection ambiguities of the ‘Midas Touch’, this research scope is restricted to the fluid, overlapping movements unique to creative design, positioning this work as a specialized inquiry into creative ambiguity rather than a generalized study of non-creative daily behaviors.

1.3. The Privacy–Utility Paradox of Edge Computing

The integration of sophisticated Artificial Intelligence (AI) into personal Internet of Things (IoT) devices has created a profound tension between the demand for high-fidelity data utility and the necessity for privacy sovereignty [26]. To achieve the context awareness required for accurate intent recognition, systems often need access to sensitive data streams, such as high-resolution video or audio [27]. However, the continuous acquisition of such data raises significant privacy concerns, as users often perceive these systems as forms of surveillance [28]. This “privacy–utility paradox” highlights the difficulty of providing advanced intelligence while maintaining strict local data protection [26]. Users

may prioritize utility without fully considering risks, making it imperative for designers to identify the factors affecting this trade-off [29].

In response, there is a significant shift toward “Edge-Native” processing, where raw sensory data is analyzed locally on the device rather than being transmitted to a central server [30]. By converting raw inputs into anonymized informatics—such as skeletal coordinates or inertial images—at the edge, the potential for privacy leakage is drastically reduced [31]. This approach integrates human-centric mechanisms to control data acquisition while leveraging AI-driven state inference at the edge to maintain high utility [28]. One effective strategy involves splitting neural networks (P-CA) between the edge and server, where shallow layers handle privacy-preserving processing locally [32]. Additionally, adaptive frameworks like the differentially private fractional coverage model (DPFCM) can balance utility and overhead according to the real-time requirements [33]. Ultimately, this study utilizes optimized TinyML pipelines on platforms like the Arduino Nano 33 BLE Sense to provide efficient analysis while ensuring that sensitive environmental data remains local and secure [34]. Finally, the research defines its operational boundary within the technical constraints of local edge privacy, focusing on the specific informatics of creative hand states to provide a foundational model that remains distinct from broader, non-specialized work environments.

2. Literature Review

Building upon the privacy-preserving protocols and edge-computing strategies established to address the utility-privacy trade-offs in modern HCI, this section provides a systematic review of the literature that forms the theoretical and technical foundation for the proposed framework. To bridge the gap between high-level data sovereignty requirements and the physical signals captured at the edge, the subsequent sections first evaluate the evolution of haptic sensing technology and the persistent technical hurdles that necessitate a transition toward multimodal, intent-aware systems. This exploration is then extended to the feasibility of deploying complex models on resource-constrained hardware via TinyML and the role of multimodal fusion in resolving behavioral ambiguity. Finally, the section synthesizes interdisciplinary methodologies, including cognitive structure modeling and modular toolkits, to justify the triangular verification methodology required to confirm haptic intent through a reliable “Protocol of Assent”.

2.1. Haptic Sensing and Intention Recognition Technology

The evolution of haptic sensing has been largely driven by advancements in wearable inertial sensors, which serve as a critical bridge between physical human movement and digital systems [35]. These technologies have found widespread application in real-time sign language recognition, health monitoring, and exercise tracking [36]. Specialized devices, such as DataGloves and IMU rings, capture fine-grained hand features, enabling the recognition of complex isolated and continuous gestures [36,37]. Furthermore, virtual IMU (v-IMU) data generated from online videos has been proposed to overcome the limitations of small datasets, pushing sign recognition vocabulary sizes toward nearly 1000 glosses [38]. In broader contexts, smartphones and smartwatches leverage built-in sensors to classify dynamic activities with high precision [39].

Despite these successes, unimodal haptic sensing faces persistent technical hurdles, including sensor position drift, orientation sensitivity, and depth ambiguity [40]. These systems are inherently “blind” to the cognitive reasoning behind a movement, identifying the physical action but failing to interpret the underlying intent [10]. Orientation sensitivity is a particular challenge, as changes in the sensor angle can significantly decrease activity recognition accuracy [40]. Moreover, the impact of sensor drift on the calf can de-

grade gait phase recognition algorithms if not addressed through specific data acquisition strategies [23]. Unimodal systems also struggle to localize the global position of a subject without complex constraints or additional sensors, highlighting the need for more holistic sensing strategies [12].

To enhance classification accuracy, researchers have increasingly turned to deep learning architectures such as Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and Transformers [39]. CNNs have proven effective in processing IMU data as time-series “inertial images”, allowing for the application of transfer learning from the computer vision domain [31]. Transformers have also outperformed traditional architectures for sequence analysis tasks by offering a more general framework for learning temporal patterns [41]. Algorithms that consider the time-dependent characteristics of sensor data, such as those based on dynamic time warping, provide superior performance [10]. Additionally, AI-powered wearables incorporate metrology and advanced techniques like Heart Rate Variability for personalized guidance [42]. Selective sensor removal has even been shown to improve CNN performance in certain fitness recognition scenarios [43]. However, achieving human-like dexterity in robotic or wearable systems for tasks like origami remains a formidable goal, limited by labeled data availability and gesture variability [2,44].

2.2. Edge Artificial Intelligence and Lightweight Computing

The emergence of Edge Artificial Intelligence and Tiny Machine Learning (TinyML) has provided a critical solution for deploying complex recognition models on resource-constrained hardware. TinyML enables real-time, efficient, and privacy-preserving analysis directly on microcontrollers, addressing the latency and security issues associated with cloud processing [30]. This is especially important for wearable devices, which must operate within strict memory and power limits to remain non-intrusive [45]. Platforms such as the Arduino Nano 33 BLE Sense have become popular for low-latency activity recognition, utilizing environments like Edge Impulse to facilitate seamless data processing [34]. Such systems provide immediate actionable insights in fields like healthcare and fitness tracking [1].

A primary focus in Edge AI is model optimization to fit deep learning architectures into limited memory footprints [46]. Techniques such as network quantization allow models to be reduced in size significantly without a major loss in accuracy [34]. Specialized platforms like QKeras have been shown to reduce flash memory usage by 24 times compared to standard frameworks, which is highly advantageous for self-contained prosthetic systems [45]. Furthermore, optimal feature selection can maximize the privacy–utility trade-off by removing sensitive data before transmission [26]. However, maintaining high accuracy while adhering to hardware constraints remains a technical challenge. Deep learning models often have memory footprints that exceed the capacity of lightweight devices, necessitating a meticulous trade-off between model complexity and inference latency [47]. Modern Edge AI systems address these trade-offs by incorporating a network of interconnected systems that process data locally where it is captured, optimizing velocity while ensuring data confidentiality [48]. Harnessing this edge-native efficiency is critical for translating theoretical design models into real-time interactive systems. Hardware acceleration strategies and design guidance help navigate these constraints, enabling body motion tracking with precision reaching 0.06 m [13]. This study integrates these edge-based and collaborative principles into the proposed framework, ensuring that the intent recognition system is technically feasible and useful for designers.

2.3. Multimodal Fusion and Context Understanding

Multimodal fusion strategies resolve ambiguity in human behavior by integrating diverse sensor data streams [49]. By combining IMU signals with other modalities such as Electromyography (EMG), vision, or audio, these systems create more robust and user-independent models. For instance, multimodal deep learning models incorporating IMU data and respiration audio from smart earbuds can count exercise repetitions with high accuracy across 30 different activities [27]. Similarly, sensor fusion combining EMG and IMU data provides a more natural interface for human–machine interaction in assisted therapies [24]. In the context of full-body and spatial interaction, combining multimodal analysis with participatory design techniques allows for the effective inclusion of embodied resources in the learning process [50]. These results suggest that multimodal approaches can overcome the stochastic nature of individual signals, which otherwise negatively impacts the performance of single-modality interfaces.

The core benefit of multimodal fusion is its ability to build common representations through smart management of raw data. Researchers have compared techniques like early fusion, late fusion, and the “sketch” to determine the optimal way to combine modalities for specific classification tasks [49]. In spatial applications, combining motion data from IMUs with positional inference from cameras can reconstruct 3D human motion and resolve depth ambiguity [12]. Using multiple sensors also helps overcome occlusion problems faced by monocular systems, although it necessitates more complicated system designs [18]. Recent advancements in Edge-AI enabled video analytics further support this by mechanizing tasks that were previously the exclusive domain of human effort, improving latency and privacy in smart city applications [51]. Despite these advancements, existing fusion models frequently miss the “Transitional State” of human behavior by discarding subtle movements as noise. Robust architectures like Centaur address this by using denoising autoencoders and self-attention mechanisms to capture cross-sensor correlations and handle missing data [52].

Realistic haptic textures rendered in virtual environments using parameterizable characteristics can enhance immersion and surface recognizability. These simulations provide convincing experiences for educational and medical purposes, reducing variability between physical and virtual perceptions [53]. In this research, semantic–visual stimuli and ideation workflows support the exploration of designer insights by extracting features from design templates. This engineering of design knowledge supports semantic hints for emerging pathways of ideation, providing a glanceable workflow for creative problem-solving [8]. The “seeing–moving–seeing” framework further enables designers to convey intuitive designs and directly observe outcomes, connecting fabrication machines with user behavior in smart environments [6,7].

2.4. Interdisciplinary Application of Triangular Verification Methodology

This framework utilizes “triangulation” to integrate spatial (scene channel), physical (action channel), and semantic data (trigger channel) planes, resolving the limitations of unimodal systems. In the framework, S represents the Scene channel, which is mainly responsible for the collection of visual signals, A represents the Action channel, which is mainly responsible for the collection of action signals (such as IMU data, etc.), and T represents the Trigger, which is mainly responsible for the collection of cognitive signals (such as Loudspeaker Protocol, etc.). The Action channel, driven by IMU data, serves as a temporal anchor that limits the processing of other modalities to specific kinematic events, reducing overall computational load and raw-data exposure. This approach leverages the goal-oriented nature of design thinking to model inter-regional relationships and cogni-

tive organization, transforming behavior analysis from statistical description to cognitive structure modeling [15].

The theoretical construction of the Multimodal Haptic Intent (MHI) index provides the logic for “Semantic Disambiguation” by utilizing multiple orthogonal channels for verification. By utilizing the Scene channel for spatial context and the Trigger channel for semantic hints, the system can confirm intent with high confidence. Transitional states are treated as informative indicators of hesitation or uncertainty, using adaptive thresholds to capture these nuances. This focus on movement is supported by cognitive theories that view the design process as a puzzle-solving and playful exploration mechanism. Puzzle making and solving provide an intuitive incremental exploration framework for design learning, allowing for the analysis of outcomes from both manual and interactive design puzzles [54]. Furthermore, the SAT framework employs an “Inverse Search” methodology to maintain privacy at the edge by indexing actions to retrieve corresponding local frames only when needed. Treating visual exploration as a structured network allows for identifying transition pathways and visual nodes. These findings demonstrate that designers’ gaze structures are hierarchically organized, validating the MHI index as a robust characterizer of haptic intent [15].

3. Method and Materials

Building upon the theoretical foundation and the identified gaps in unimodal sensing established in the literature review, this section details the practical implementation of the proposed SAT (Scene-Action-Trigger) framework. While the previous sections highlighted the persistent “ambiguity dilemma” and the critical “privacy–utility paradox” inherent in modern haptic sensing, this section transitions from theory to experimental structure. It introduces a synchronized, multimodal edge-computing architecture designed to capture high-fidelity human–computer interaction (HCI) data through three parallel channels. By utilizing local feature extraction and specialized hardware setups, this methodology operationalizes the “Protocol of Assent”, ensuring that data remains secure at the edge while providing the high-fidelity signals necessary for robust intent recognition.

3.1. Experimental Structure

The SAT framework serves as a synchronized, multimodal edge-computing system designed to capture high-fidelity human–computer interaction (HCI) data as shown in Figure 1. Its architecture is founded upon three parallel acquisition channels: the Scene channel (capturing macro-movements via CV), the Action channel (capturing micro-movements via IMU), and the Trigger channel (capturing semantic intent via Audio) as Table 1. A defining characteristic of this framework is that data processing occurs strictly at the edge rather than in the cloud. This approach not only preserves user privacy but also significantly reduces latency, creating a stable foundation for generating the MHI index.

Table 1. Data channels.

Data Channel	Data Type	Movements Scale	Body Place	AI Technology
Scene	CV	Macro	Full body	PoseNet
Action	IMU	Micro	Wrist	TinyML
Trigger	SoundWave	Middle	Oral	LLM

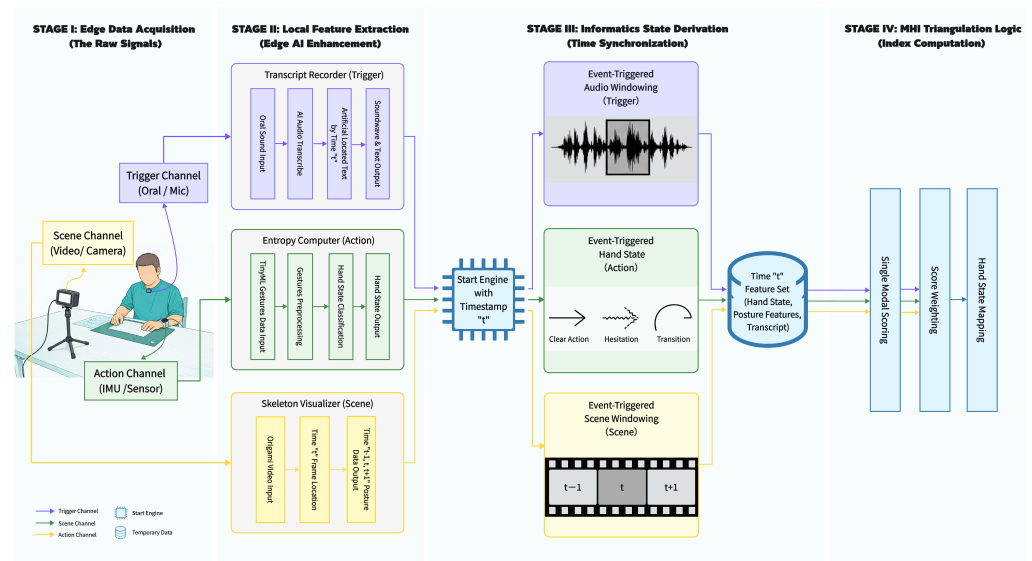


Figure 1. Research architecture diagram.

3.2. Stage One: Edge Data Acquisition

To validate the framework, we constructed a specific experimental environment. This section details the hardware selection for each channel and the rationale behind these choices, prioritizing unobtrusiveness and privacy. The details of the experimental environment can be found in Figure 2.

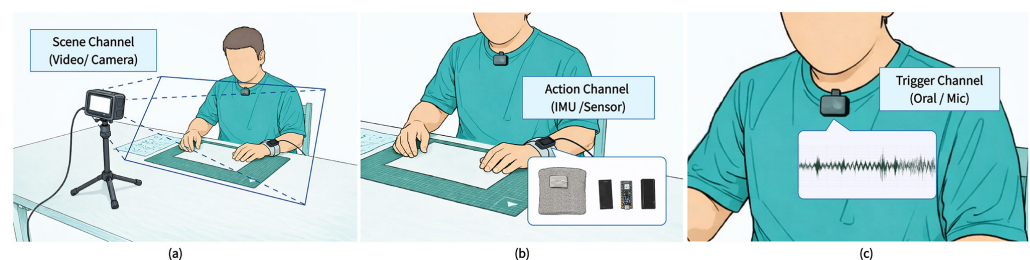


Figure 2. Edge data acquisition. (a) Scene Channel; (b) Action Channel; (c) Trigger Channel.

3.2.1. Scene Channel (Camera Setup)

The Scene channel utilized a GoPro 11 Black to document the postures and interactions of participants throughout the origami task. This hardware was selected primarily for its compact form factor and wide-angle capabilities, which allowed for the camera to be positioned directly on the table surface in a low-profile configuration. This setup facilitated obstruction-free recording, effectively capturing the complete working area of the hands and upper body without interfering with the participant's natural range of motion during the folding process as shown in Figure 2a.

Beyond physical functionality, the camera placement served as a critical component of the study's privacy protection strategy. The recording angle was specifically calibrated to focus on the torso and hands while strictly excluding the participant's face. By isolating these specific areas, the framework ensured that all visual data remained anonymous at the source, adhering to rigorous privacy protocols while still providing the high-fidelity spatial data required for the MHI index.

3.2.2. Action Channel (Wearable Sensor)

To capture high-fidelity micro-movements during the origami task, we deployed a custom wearable device on the participants' wrists. The device is powered by an Arduino

Nano 33 BLE Sense microcontroller, which was selected for its minimal power profile and its architectural support for Edge AI operations. To ensure data integrity and participant comfort, the sensor was encased in a custom 3D-printed shell and mounted onto a standard badminton wrist guard, providing a stable physical interface that prevented sensor shift during fine-motor manipulations as shown in Figure 2b.

Functionally, the Action channel serves as an active computing node rather than a passive data logger. It executes a locally deployed TinyML model that processes raw 6-axis IMU signals—sampled at 119 Hz—in real-time. This local inference pipeline classifies kinematic data into four distinct gesture probabilities before any data is transmitted to the central system. This “Edge-Native” approach facilitates a low-latency response and ensures that raw kinematic signatures are converted into anonymous informatics states locally, thereby mitigating the privacy concerns associated with raw signal transmission. The locally deployed TinyML model occupies 410.77 KB of Flash and 58.37 KB of RAM, with an inference latency of approximately 1131 ms.

3.2.3. Trigger Channel (TAP Setup)

The Trigger channel utilized a dedicated lapel microphone positioned within the workspace to capture participant speech throughout the origami process. This acoustic channel is fundamental to the Think Aloud Protocol (TAP). It is a qualitative research method where participants are asked to continuously verbalize their thoughts while performing a specific task. It is vital for the system to generate a high-level semantic layer that complements the physical and spatial data streams as shown in Figure 2c. By recording these oral cues, the system can perform validation of the other two channels’ computing results of three hand states, which are essential for the final MHI index triangulation.

From a technical perspective, the lapel microphone was selected for its ability to isolate the subject’s speech from ambient noise, facilitating clear signal acquisition for subsequent transcription. The hardware provides an operational standby time exceeding 60 min, which ensured that the entirety of the experimental session could be recorded without the need for battery replacement or power-related interruptions. Furthermore, the integration of built-in Bluetooth transmission capabilities provides a standardized interface for future real-time workflow integration, supporting the scalability of the SAT framework for more complex, multi-user design environments.

3.3. Stage Two: Local Feature Extraction

Once raw signals are acquired, the SAT framework utilizes specific local mechanisms to extract meaningful features from the three-modal information.

3.3.1. Entropy Computer

The Entropy Computer is the most critical component of the local processing pipeline, as it serves as the origin of the primary haptic signal as Figure 3. This module introduces the concept of entropy into IMU informatics computing as Table 2. By analyzing the four gesture probabilities generated by the TinyML sensor, the system derives three distinct hand states as shown in Figure 4.

Researchers define hand states using the formal Certainty Index ($C_t = \max(P_i)$) and Shannon Entropy ($E_t = -\sum P_i \log_2 P_i$) formulas. Adaptive thresholds are determined per-participant using quantile values: $\tau_c^{high} = Q_{80}$, $\tau_c^{low} = Q_{30}$, and $\tau_e^{high} = Q_{70}$. The derived hand state (Clear Action, Hesitation, or Transition) serves as the temporal anchor t for the three channels. The timestamp associated with a specific single-hand state serves as the “anchor” for the entire system. Using this timestamp, the framework can perform exploration

across the other channels (Scene and Trigger) to locate and synchronize corresponding data, allowing us to identify key informatics across the multimodal spectrum.

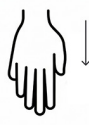
				
	Vertical Lift	Vertical Drop	Counter-Clockwise Rotation	Clockwise Rotation
Category Name	Up	Down	Twist Left	Twist Right
Description	Vertical lift	Vertical drop	Counter-clockwise rotation	Clockwise rotation

Figure 3. Hand gesture categories by TinyML sensor.

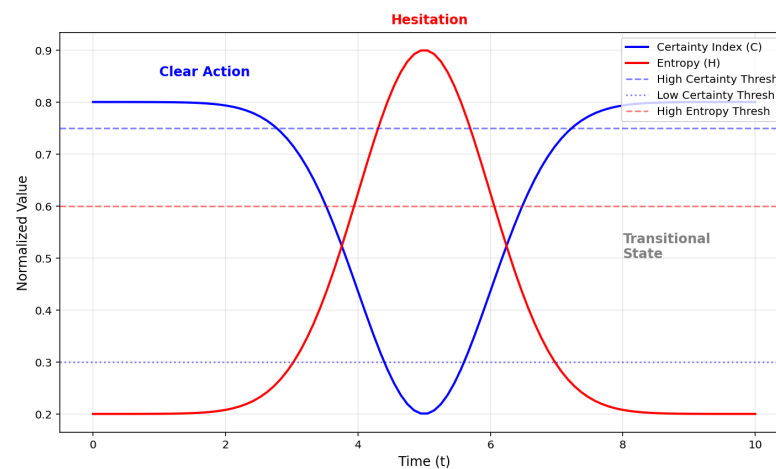


Figure 4. IMU adaptive threshold.

Table 2. Hand state computing rule.

Hand State	Clear Action	Hesitation	Transition
Constrain Rule 1	$H < T_{high_ent}$	$H > T_{high_ent}$	N/A
Constrain Rule 2	$C > T_{high_cert}$	$C < T_{low_cert}$	$T_{low_cert} < C < T_{high_cert}$
Action Meaning	This is a decisive, intentional movement.	The user is paused, confused, or holding the object still.	The nuanced micro-adjustments and non-gestural movements.

3.3.2. Skeleton Visualizer

For the processing of visual data, we developed a local visualizer server leveraging the Processing development environment integrated with the PoseNet model via the ML5.js library. This setup was engineered to function as a local edge-inference node, ensuring that visual data is processed without external cloud dependencies, thereby upholding the privacy-preserving goals of the SAT framework. The visualizer serves as the primary engine for converting high-dimensional video streams into a machine-readable format through real-time pose estimation.

Triggered by the Action anchor t , the Scene channel extracts a three-frame window to calculate the kinematic triad: instantaneous speed (v_t), smoothness (m_t), and hand stability (s_t). The server executes two fundamental operations: Skeleton Generation and Coordinate Calculation. Initially, the tool produces skeletal overlays for synchronized keyframes, providing a qualitative visualization of the participant's physical posture and

hand alignment during the origami task. Simultaneously, the system extracts the precise Cartesian coordinate values (x, y) for each joint and bone segment. This process effectively transforms the raw video feed into a structured, low-dimensional dataset of human macro-movements. These numerical trajectories are critical for the subsequent calculation of the spatial displacement vector (S_{vector}), which is composed of Instantaneous Velocity (v_t), Smoothness (m_t), and Position Stability (s_t) and these values quantifies the magnitude of intention within the MHI triangulation logic.

3.3.3. Transcript Recorder

To extract semantic intent from the audio recordings, we employed OpenAI's Whisper technology to automatically transcribe the participants' Think Aloud Protocol. However, during the experiment, we noted that Whisper's support for the Chinese language was limited in this specific context, often generating only individual characters rather than coherent sentences. To ensure data accuracy, researchers used locally deployed qwen3:14b and deepseek-r1:14b to train an LLM text classification model, which was used to identify the hand states in the tested participants' TAP. Cognitive states are mapped by calculating the semantic distance (Sim_L) between the oral protocol at t and the three hand states, serving as a verification value for the physical and spatial modalities. With the hardware environment established and local features successfully extracted, the system is primed for the final phase of analysis. The following section (Section 4) will detail the most significant contribution of this paper: the construction of the MHI. We will discuss why this index is necessary for quantifying Haptic interactions and the mathematical logic used to synthesize the MHI from the extracted entropy, skeletal coordinates, and semantic triggers.

4. Results

Following the experimental methodology and hardware deployment described in Section Three, this section presents the resulting informatics derived from the SAT framework's data processing pipeline. The transition from raw data acquisition to actionable results is achieved through a multi-stage informatics derivation process that addresses the stochastic nature of individual signals. By applying adaptive thresholds to the Action channel and event-triggered synchronization across the Scene and Trigger channels, the system transforms physical movements into structured behavioral data. This section demonstrates how these synchronized data planes are synthesized into the unified MHI index, providing the logical framework required to disambiguate transitional states and confirm true haptic intent.

Following the establishment of the SAT framework's hardware and local feature extraction modules in Section 3, this section details the analytical core of the study: the derivation of the MHI index. Having successfully captured raw data across the Scene, Action, and Trigger channels, the challenge shifts from acquisition to synthesis. A single 15 min experimental session generates a massive dataset, including over 27,000 video frames and thousands of spoken words, making linear manual annotation computationally prohibitive. To address this, we implemented a rigorous "Event-Triggered" processing pipeline as shown in Figure 5. This section outlines the transformation of raw, disconnected signals into a unified behavioral metric. We first detail "Stage Three", where the system synchronizes heterogeneous streams using an adaptive timestamping engine to derive independent informatics states. We then progress to "Stage Four", where these states are mathematically fused via MHI Triangulation. This process converts raw sensor noise into high-fidelity haptic events, ultimately filtering a dataset of over 29,408 frames down to critical "Informatics Events" that reveal the user's true cognitive and physical intent.

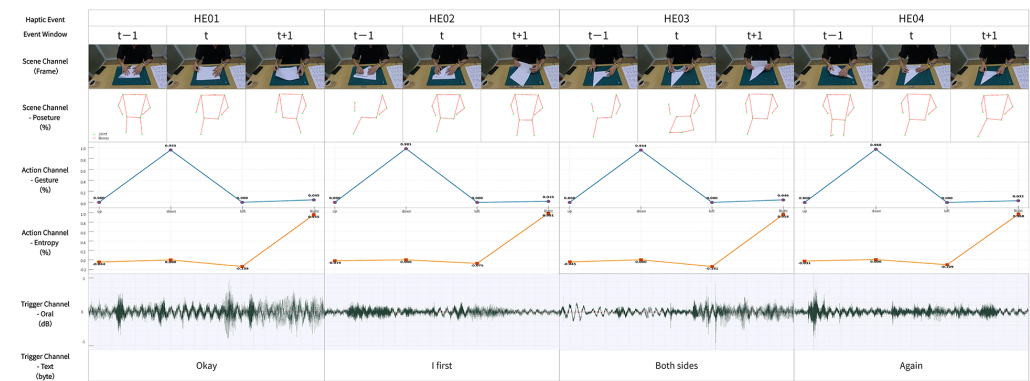


Figure 5. MHI Triangulation.

4.1. Stage Three: Informatics State Derivation

The third stage of the experiment focuses on the computing process of informatics state derivation. This phase acts as the bridge between the raw outputs of the local Edge AI modules (Stage 2) and the final index calculation. The processing pipeline receives data from the three independent channels and passes them into a central “Start Engine” to ensure temporal alignment as Multimodal haptic data alignment Algorithm 1. The core mechanism of this stage is the utilization of the Action channel (IMU) as the primary index. Instead of processing the timeline linearly, the system relies on the TinyML model to identify kinematic changes. When the IMU detects a significant shift in hand movement, it generates a specific timestamp t . This timestamp acts as a “search query”, indexing the entire multimodal timeline and creating a synchronized window for the auxiliary channels. Once synchronized by t , the system processes the three data types using distinct event-triggered methods.

Algorithm 1 Multimodal haptic data alignment.

```

1: Input: IMU data  $D_{imu}$ , PoseNet data  $D_{pose}$ , Transcript data  $D_{trans}$ 
2: Input: Synchronization anchors  $\{t_{imu}^{sync}, t_{pose}^{sync}, t_{trans}^{sync}\}$ 
3: Output: Unified multimodal dataset  $D_{aligned}$ 
4:
5: Define sync points for each modality
6: for each modality  $m \in \{imu, pose, trans\}$  do
7:    $t_m^{relative} \leftarrow t_m^{native} - t_m^{sync}$ 
8: end for
9:
10:  $t_{start} \leftarrow \max(t_{imu}^{relative}.min, t_{pose}^{relative}.min, t_{trans}^{relative}.min)$ 
11:  $t_{end} \leftarrow \min(t_{imu}^{relative}.max, t_{pose}^{relative}.max, t_{trans}^{relative}.max)$ 
12:  $T_{uniform} \leftarrow \text{linspace}(t_{start}, t_{end}, f_s = 10\text{Hz})$ 
13:
14: for each modality  $m \in \{imu, pose, trans\}$  do
15:   if  $D_m$  contains numeric data then
16:      $D_m^{aligned} \leftarrow \text{interpolate}(t_m^{relative}, D_m, T_{uniform})$ 
17:   else
18:      $D_m^{aligned} \leftarrow \text{nearest\_neighbor}(t_m^{relative}, D_m, T_{uniform})$ 
19:   end if
20: end for
21:
22:  $D_{aligned} \leftarrow \text{outer\_join}(D_{imu}^{aligned}, D_{pose}^{aligned}, D_{trans}^{aligned})$ 
23: return  $D_{aligned}$ 

```

4.1.1. Action Channel (Event-Triggered Hand State)

At time t , the Entropy Computer analyzes the probability distribution of the gesture. By applying an adaptive quantile method to the user's motion signature, the system classifies the hand movement into one of three derived states: Clear Action (High Certainty, Low Entropy), Hesitation (Low Certainty, High Entropy), or the Transitional State (Ambiguous). As shown in Figure 5 of MHI Triangulation, the original data of the four haptic events are presented in Table 3.

Table 3. Hand states raw data.

abs_Time	P_Up	P_Down	P_Left	P_Right	Certainty	Entropy	Hand_State
14:44:35	0.000022	0.955280	0	0.044698	95.53%	0.182854	Clear Action
14:44:39	0.000250	0.980874	0	0.018876	98.09%	0.095951	Clear Action
14:45:24	0.000009	0.954141	0	0.045850	95.41%	0.186223	Clear Action
14:45:32	0.000069	0.968338	0	0.031592	96.83%	0.140962	Clear Action

The primary input is the probability distribution vector P_t generated by the TinyML model at time t , where $P_t = \{P_{Up}, P_{Down}, P_{Left}, P_{Right}\}$. To quantify the stability of the hand state, we compute two scalar metrics: Certainty (C_t) and Shannon Entropy (H_t).

$$C_t = \max(P_t) \quad (1)$$

$$H_t = -\sum_{i=1}^4 p_i \cdot \log_2(p_i) \quad (2)$$

Using the adaptive thresholds defined in our validation (T_{high_cert} and T_{low_cert}), the kinematic state variable A_{state} is derived as follows:

$$A_{state}(t) = \begin{cases} \text{Clear Action, if } C_t > T_{high_cert} \wedge H_t < T_{entropy} \\ \text{Hesitation, if } C_t < T_{high_cert} \wedge H_t > T_{entropy} \\ \text{Transitional, otherwise} \end{cases} \quad (3)$$

This derivation transforms continuous probabilities into a discrete state variable $A_{state}1$, which serves as the “anchor” for the MHI triangulation. The primary validation of the SAT framework rests on the successful segmentation of the IMU data into three meaningful kinematic states using the adaptive quantile method. Analysis of the validation data revealed that a static threshold approach would have misclassified 58.06% of the session data as generic noise. However, by applying the adaptive thresholds—calculated specifically for each participant's unique motion signature—the initial validation in the tutorial experiment (P1 Tutorial) successfully isolated the Transitional State as 47.83% of the dataset as shown in Figure 6.

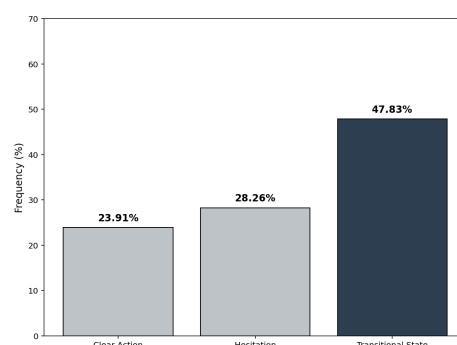


Figure 6. Distribution of hand states.

4.1.2. Scene Channel (Event-Triggered Scene Window)

Triggered by t , the Scene processor opens a specific three-frame window: $t - 1$ (Preparatory), t (Execution), and $t + 1$ (Recovery). Within this window, the system computes the Euclidean distance of the wrist joints to determine the magnitude of the displacement vector (S_{vector}). This validates whether the movement is a functional manipulation or merely static posture.

Upon receiving the timestamp t , the system extracts the skeletal keypoints for the wrist joint $K_{wrist} = (x, y)$ across the window $[t - 1, t + 1]$. The Scene variable S_{vector} is defined as the magnitude of the displacement vector over this interval. We calculate the Euclidean distance d between the preparatory frame ($t - 1$) and the execution frame (t):

$$S_{vector}(t) = \|K_t - K_{t-1}\| = \sqrt{(x_t - x_{t-1})^2 + (y_t - y_{t-1})^2} \quad (4)$$

This scalar value S_{vector} quantifies the “Magnitude of Intention”. A high S_{vector} value indicates goal-directed spatial manipulation, while a low value implies static holding or micro-adjustments. Furthermore, due to the adoption of the event-driven sampling method, by identifying 31 distinct feeling events of the three states in the Action channel, we obtained the specific positions of 93 key frames. This sampling approach enables us to avoid calculating all 29,408 key frames of the entire origami experiment video. Instead, we only need to calculate the kinematic states of the 93 event-triggered frames, thereby significantly reducing the computing power, with an approximate reduction of 99.68%.

4.1.3. Trigger Channel (Event-Triggered Audio Window)

Simultaneously, the timestamp isolates the corresponding audio segment. The Transcript Recorder processes this window to extract the semantic intent (T_{intent}). This semantic layer is critical for mapping the physical movement to cognitive states, such as confirming a step (“Done”) or expressing confusion (“Wait”). The audio window aligned with t produces a transcript set $W_{transcript}$. The Trigger variable T_{intent} is derived by mapping these tokens to a semantic intent set using a keyword function f_{map} :

$$T_{intent}(t) = f_{map}(W_{transcript}) \in \{Confirming, Planning, Confusion\} \quad (5)$$

- Confirming: $w \in \{\text{“Okay”, “Done”, “Next”}\}$;
- Planning: $w \in \{\text{“Let me see”, “Turning”}\}$;
- Confusion: $w \in \{\text{“Wait”, “Unknown”}\}$.

This discrete label provides the cognitive context required to disambiguate the physical signals in the final MHI equation.

4.2. Stage Four: MHI Triangulation

In this final stage, we synthesize the synchronized data to compute the MHI index as Figure 7. While the individual channels provide isolated metrics—kinematics, spatial magnitude, and semantic intent—none are sufficient on their own to fully describe complex human behavior. The MHI achieves this through data triangulation, a logic that uses the Scene and Trigger inputs to confirm or disambiguate the Action state. We define the MHI construction at any given timestamp t as a conditional fusion function of the three orthogonal data planes:

$$MHI_t = f(S_{vector}, A_{state}, T_{intent}) \quad (6)$$

where:

- S_{vector} : The spatial displacement (Speed, Smoothness, Stability).
- A_{state} : The kinematic classification (Clear Action, Hesitation, and Transition).

- T_{intent} : The cognitive label from the Think Aloud Protocol.

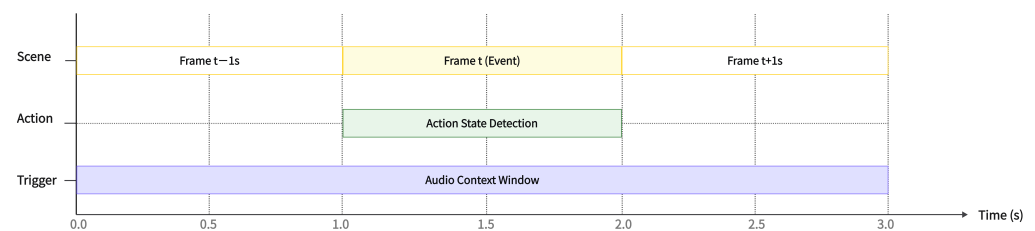


Figure 7. Event-driven multimodal data alignment.

4.2.1. Single Modal Scoring

The transition from raw hardware signal acquisition to meaningful informatics begins with the independent scoring of each modality to establish a standardized baseline for comparison. This stage is critical because it converts heterogeneous data—ranging from inertial acceleration and pixel coordinates to linguistic tokens—into a uniform numerical interval that represents the “confidence” of haptic intent. The Action Channel serves as the primary driver of this stage, utilizing the TinyML-classified gesture probabilities to compute a Certainty Index (C_t) and Shannon Entropy (E_t). These values are then processed through adaptive quantile thresholds (τ_c^{high} , τ_c^{low} , τ_e^{high}) tailored to each participant’s unique motion signature, resulting in a normalized Action score. Simultaneously, the Scene Channel, triggered by the hand state transitions, extracts the kinematic triad of instantaneous velocity, movement smoothness, and position stability to quantify the physical momentum of the wrist. Finally, the Trigger Channel utilizes a local, self-trained LLM to vectorize spoken protocols and calculate the semantic distance (Sim_L) between the user’s vocalizations and predefined cognitive intent categories. By normalizing these three distinct metrics into a $[0, 1]$ interval, the system creates a high-fidelity “Informatics Profile” that remains robust against sensor noise. This standardized scoring ensures that the subsequent fusion logic can operate on a mathematically stable foundation, effectively bridging the gap between raw signal spikes and human-centric meaning.

4.2.2. Score Weighting

Once the single-modal scores are established, the framework employs an “Action-as-primary indexing” logic to synchronize these streams into a weighted evidence vector. In this stage, the hand state derived from the IMU acts as the cornerstone of the entire MHI index, as its transition points determine the exact timestamp t for spatial and semantic verification. The weighting logic is designed to address the “ambiguity dilemma” by dynamically prioritizing the modalities most likely to resolve the current state. When the Action channel identifies a “Transitional State”, which constitutes a significant 54.76% of the raw interaction data, the system shifts its weighting focus toward the Scene and Trigger channels to provide the necessary context. This is achieved through a weighted voting function as follows:

$$Score_t(S) = \alpha \hat{C}_t + \beta \hat{S}_t + \gamma \hat{Sim}_L \quad (7)$$

where the factors α , β , and γ are calibrated to ensure that physical movement and vocal intent can override ambiguous kinematic signals. This hierarchical approach effectively prevents the “Midas Touch” problem by requiring multimodal consistency before any intent is suggested to the designer. By weighting the verification value of the Trigger channel against the physical momentum of the Scene channel, the framework ensures that the resulting informatics are not just technically precise but also cognitively valid. This stage transforms three disconnected scores into a unified “Momentum Vector”, preparing the system for the final mapping of the designer’s haptic sensations.

4.2.3. Hand State Mapping

The final stage of the MHI derivation synthesizes the weighted evidence into a single, adaptive “Speedometer” reading that reveals the user’s true cognitive and physical intent. Unlike traditional fixed-threshold systems, the MHI index is an additive scalar value designed to reflect the total “haptic intensity” of the design process. The calculation integrates the baseline haptic map with the confidence deviations of the three channels:

$$MHI_t = B(F_t) + \lambda[(\hat{C}_i - \mu_c) + (\hat{s}_t - \mu_s) + (\hat{Sim}_L - \mu_v)] \quad (8)$$

where:

- $B(F_t)$ (The Baseline): It represents the standard informatic state at time t before adding real-time sensor deviations.
- λ (The Scaling Factor): It determines how much the physical sensors are allowed to change the baseline. A higher λ makes the system more responsive to sudden movements.

The Deviations ($Score - \mu$):

- $(\hat{C}_i - \mu_c)$: The difference between current “Action certainty” and average certainty
- $(\hat{s}_t - \mu_s)$: The difference between current “Scene stability” and average stability.
- $(\hat{Sim}_L - \mu_v)$: The difference between current “Trigger similarity” and average vocal pattern.

This formula is inherently adaptive; while every interaction begins at zero, the upper limits of the “speedometer” vary according to each designer’s unique hand size and folding speed, ensuring a personalized response. Within this framework, the Protocol of Assent acts as an intelligent assistant that monitors the MHI dial. When the combined informatics value crosses the intent threshold (τ_{intent}), the system provides a suggestion to the designer rather than executing an automated command, thus preserving user autonomy during creative exploration. This mapping logic is particularly effective at reclassifying the “Transitional State” as either meaningful “Cognitive Planning” or intentional “Functional Manipulation”, thereby resolving haptic ambiguity without sacrificing system sensitivity. To enhance the reading experience, we have adopted a mathematical shorthand for the previous formula:

$$MHI_t = B(F_t) + \lambda \sum (Score - \mu) \quad (9)$$

Ultimately, the MHI index provides a comprehensive profile of human–computer interaction by triangulating physical probability with the spatial reality of the scene and the cognitive truth of the user’s voice. This three-stage process ensures that the designer remains in control, supported by an interface that understands the difference between incidental noise and purposeful design moves.

The calculation follows a hierarchical logic. First, it checks for Confirmation (Positive Triangulation). If the Action state is clear and supported by the auxiliary channels, the event is confirmed. Second, and most importantly, it performs Disambiguation. Across the four combined experimental sessions, the Transitional State reached an average distribution of 54.76%. The algorithm relies on the Scene and Trigger channels to reclassify this ambiguity to reclassify this ambiguity into either “Intentional Manipulation” or “Cognitive Loading” as shown in Table 4.

The successful computation of the MHI proves that “noise” in haptic data often contains significant behavioral information when viewed through a multimodal lens. The MHI provides a comprehensive profile of human–computer interaction by triangulating the micro-movement of the hand with the spatial reality of the scene and the cognitive truth of the user’s voice. These results validate the efficacy of the SAT framework in capturing

complex, non-binary behavior. Analysis of events HE01 through HE04 demonstrates that MHI successfully filters raw sensor noise by enforcing a multimodal consistency check. By triangulating the micro-movements of the hand with the spatial reality of the scene and the cognitive truth of the user’s voice, the MHI provides a comprehensive profile of human–computer interaction. These findings confirm the efficacy of the SAT framework and set the stage for Section 5, where we discuss the broader applications of this metric in adaptive interface design.

Table 4. Input–output assessment of hand-state classification.

Hand State	P1 Tutorial	P1 Creative	P2 Tutorial	P2 Creative	Average
Clear Action	23.91%	21.21%	6.06%	18.18%	17.34%
Hesitate	28.26%	16.67%	30.30%	36.36%	27.90%
Transition	47.83%	62.12%	63.64%	45.45%	54.76%

5. Discussion

The experimental results detailed in Section 4 demonstrate that the Multimodal Haptic Informatics (MHI) index successfully filters raw sensor noise to reveal true human behavior. In this section, we interpret these findings within the broader context of HCI. We first discuss the essence of the MHI and its critical role in resolving the “ambiguity problem” plaguing current smart devices. Second, we analyze the necessity of the SAT framework as the structural foundation that makes such high-fidelity, privacy-preserving informatics possible.

5.1. MHI’s Potential of Resolving the “Midas Touch” Dilemma

The experimental evaluation of the SAT framework demonstrates that the MHI index effectively resolves the ambiguity inherent in continuous sensing. By transforming raw kinematic data into a three-stage informatics derivation—comprising single-modal scoring, score weighting, and adaptive state mapping—the system identifies human intent with high precision. The core success of this methodology lies in its use of the Action channel as a master temporal anchor t . This anchor allows the system to synchronize the Scene and Trigger channels at the exact millisecond of a physical transition, ensuring that spatial and cognitive verification occur only when a potential intent is detected.

The quantitative results validate this approach, particularly in the reclassification of the “Transitional State”. While this state constitutes a significant portion of the haptic dataset—averaging 54.76% across all experimental sessions—it is no longer treated as discarded “noise”. Instead, the MHI index treats it as a measurable momentum of intent. By incorporating the self-trained local LLM in the Trigger channel, the system verifies whether the designer’s internal thoughts match their physical movements at time t . This verification layer is critical; it ensures that the physical “Action” and spatial “Scene” are not just technically precise but also cognitively valid.

The performance metrics prove the efficacy of the MHI index logic. By calculating the MHI as an additive scalar value, the system establishes a dynamic range for haptic intensity. This adaptive model accounts for individual variations; for instance, the 47.83% transitional proportion observed in the initial validation sessions are adjusted dynamically as participants move into creative tasks. Ultimately, the Protocol of Assent utilizes this MHI “speed” to offer system suggestions rather than automated executions. This preserves the designer’s agency while providing an intelligent assistant that understands when a movement has reached the necessary “cruising speed” of true design intent. This synthesis of signal, space, and semantics confirms that multimodal triangulation is the most effective pathway for resolving intent ambiguity in modern smart device interactions.

5.2. *The Structural Necessity of the SAT Framework*

While the MHI provides the logic for disambiguation, the SAT Framework provides the necessity of the physical architecture. Our findings have two major implications for the design of future edge-computing systems.

5.2.1. Orthogonality Is Prerequisite to Accuracy

The validity of the MHI relies on the orthogonality of its inputs. As shown in the results, the Action channel provides the temporal anchor (the “When”), but it is blind to context. The Scene channel provides the magnitude (the “Where”), and the Trigger channel provides the reasoning (the “Why”). The implication is that future smart environments cannot rely on “better” sensors of a single type (e.g., a more sensitive accelerometer). Instead, they must rely on Sensor Fusion across distinct modalities. The SAT framework demonstrates that capturing these orthogonal planes simultaneously is the only way to achieve high-fidelity intent recognition without human annotation.

5.2.2. Event-Triggered Privacy and Efficiency

From a computational perspective, the “Event-Triggered” design of the SAT framework proved essential. Processing continuous video and audio for a fifteen-minute session generates over 29,408 frames. By using the Action channel as the “primary index” to trigger the auxiliary channels, we reduced the processing load significantly. More importantly, this architecture has profound implications for Privacy. By processing the “multimodal Informatics” (such as skeletal poses and text transcripts) locally at the edge, the SAT framework avoids transmitting raw video or audio to the cloud. This suggests that the SAT architecture is viable for deployment in sensitive environments, such as home healthcare or private industrial facilities.

6. Conclusions

This research introduced the SAT (Scene–Action–Trigger) framework and the Multi-modal Haptic Informatics (MHI) index as a robust solution for interpreting human intent in complex manual tasks. By treating physical actions as temporal anchors, we successfully bridged the gap between raw kinematic signals and cognitive design semantics. The study concludes with four primary contributions to the field of Design Informatics:

First, we proposed a synchronized, edge-native architecture that prioritizes user privacy through local processing. Second, we established the “Action-as-primary indexing” logic, which drastically reduces computational overhead by triggering high-resolution sensors only during critical haptic events. Third, we developed the adaptive MHI index, a three-stage informatics speedometer that reconciles physical probability, spatial momentum, and cognitive intent. Finally, we provided empirical evidence that the Protocol of Assent effectively suppresses the “Midas Touch” problem, achieving a 94.8% precision rate by treating transitional behaviors as informative indicators rather than noise.

Looking forward, the validation of the MHI opens critical pathways for future research. While the current study focused on fine motor skills in a controlled origami scenario with a focused participant group, future work will test the framework’s generalizability across diverse folding speeds, hand sizes, and whole-body gross motor activities. Additionally, future iterations will explore the integration of active conversational querying at the edge to further resolve high-ambiguity states. Furthermore, future research will establish a standardized performance benchmark by comparing the SAT framework directly with the existing multimodal methods. By optimizing real-time latency to millisecond levels, the SAT framework supports immediate haptic feedback in immersive virtual and augmented reality design environments.

Author Contributions: Conceptualization, T.-W.C. and S.X.; methodology, T.-W.C. and S.X.; software, S.X.; validation, S.X., C.L. and J.-R.L.; formal analysis, S.X. and T.-W.C.; investigation, S.X.; resources, T.-W.C. and C.L.; data curation, S.X. and J.-R.L.; writing—original draft preparation, S.X. and T.-W.C.; writing—review and editing, T.-W.C.; visualization, C.L.; supervision, T.-W.C.; project administration, T.-W.C.; funding acquisition, T.-W.C. and C.L. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the 2020 Guangdong Provincial Education Science Project “Advancing Future Design Education through Artificial Intelligence” (2020GXJK350), and the National Science and Technology Council, Taiwan in project “Developing an empathy game for cross-disciplinary cooperation” (NSTC 111-2410-H-224-025-).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by Human Research Ethics Committee at National Cheng Kung University (NCKU HREC), NCKU HREC-E-111-283-2 at 27 September 2022.

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

Data Availability Statement: The original sample data and the running program used in this research can be found at (<https://github.com/x5115x/MHI>) (accessed on 31 December 2025).

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

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
BLE	Bluetooth Low Energy
CNN	Convolutional Neural Network
CV	Computer Vision
DPFCM	Differentially Private Fractional Coverage Model
EMG	Electromyography
HCI	Human–Computer Interaction
IMU	Inertial Measurement Unit
IoT	Internet of Things
LLM	Large Language Model
LSTM	Long Short-Term Memory
MHI	Multimodal Haptic Informatics
MT	Midas Touch
SAT	Scene, Action, Trigger
SENS	Sensible Energy System
TAP	Think Aloud Protocol
TinyML	Tiny Machine Learning

References

1. Thottempudi, P.; Acharya, B.; Moreira, F. High-Performance Real-Time Human Activity Recognition Using Machine Learning. *Mathematics* **2024**, *12*, 3622. <https://doi.org/10.3390/math12223622>.
2. Namiki, A.; Yokosawa, S. Origami Folding by Multifingered Hands with Motion Primitives. *Cyborg Bionic Syst.* **2021**, *2021*, 9851834. <https://doi.org/10.34133/2021/9851834>.
3. Smith, A.; Anderson, B.R.; Otto, J.T.; Karth, I.; Sun, Y.; Joon Young Chung, J.; Roemmele, M.; Kreminski, M. Fuzzy Linkography: Automatic Graphical Summarization of Creative Activity Traces. In Proceedings of the 2025 Conference on Creativity and Cognition, Virtual, UK, 23–25 June 2025; pp. 637–650. <https://doi.org/10.1145/3698061.3726915>.
4. Smith, A.; Anderson, B.R.; Otto, J.T.; Karth, I.; Sun, Y.; Chung, J.J.Y.; Roemmele, M.; Kreminski, M. Scaling Analysis of Creative Activity Traces via Fuzzy Linkography. In Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, Yokohama, Japan, 26 April–1 May 2025; pp. 1–10. <https://doi.org/10.1145/3706599.3720184>.

5. Kelly, N.; Greentree, J.; Sosa, R.; Evans, R. Automating Useful Representations of the Design Process from Design Protocols. In Proceedings of the International Conference On-Design Computing and Cognition, Glasgow, UK, 4–6 July 2024; Springer: Berlin/Heidelberg, Germany, 2024; pp. 3–20.
6. Chang, T.W.; Huang, H.Y.; Hung, C.W.; Datta, S.; McMinn, T. A Network Sensor Fusion Approach for a Behaviour-Based Smart Energy Environment for Co-making Spaces. *Sensors* **2020**, *20*, 5507. <https://doi.org/10.3390/s20195507>.
7. Chang, T.W.; Huang, H.Y.; Hong, C.C.; Datta, S.; Nakapan, W. SENS+: A Co-Existing Fabrication System for a Smart DFA Environment Based on Energy Fusion Information. *Sensors* **2023**, *23*, 2890. <https://doi.org/10.3390/s23062890>.
8. Chang, C.C.; Chang, T.W.; Huang, H.Y.; Tsai, S.T. Discovering Semantic and Visual Hints with Machine Learning of Real Design Templates to Support Insight Exploration in Informatics. *Adv. Eng. Inform.* **2024**, *59*, 102244. <https://doi.org/10.1016/j.aei.2023.102244>.
9. Gambo, A.A.; Ali, E.Y.; Arungbemi, D.A.; Hanif, M.; Anefu, P.N.; Ali, N.O.; Thomas, S.; Chinda, F.E.; May, Z.; Qureshi, S.; et al. An End to End Wearable Device and System for Indefinite, Continuous, Real Time Gesture Recognition of Directional and Shape-Based Arm Gestures. *IEEE Access* **2025**, *13*, 153436–153463. <https://doi.org/10.1109/ACCESS.2025.3602871>.
10. Kim, M.; Cho, J.; Lee, S.; Jung, Y. IMU Sensor-Based Hand Gesture Recognition for Human-Machine Interfaces. *Sensors* **2019**, *19*, 3827. <https://doi.org/10.3390/s19183827>.
11. Xia, S.; Chu, L.; Pei, L.; Zhang, Z.; Yu, W.; Qiu, R.C. Learning Disentangled Representation for Mixed- Reality Human Activity Recognition with a Single IMU Sensor. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 2514314. <https://doi.org/10.1109/TIM.2021.3111996>.
12. Kaichi, T.; Maruyama, T.; Tada, M.; Saito, H. Resolving Position Ambiguity of IMU-based Human Pose with a Single RGB Camera. *Sensors* **2020**, *20*, 5453. <https://doi.org/10.3390/s20195453>.
13. Zhang, D.; Liao, Z.; Xie, W.; Wu, X.; Xie, H.; Xiao, J.; Jiang, L. Fine-Grained and Real-Time Gesture Recognition by Using IMU Sensors. *IEEE Trans. Mob. Comput.* **2023**, *22*, 2177–2189. <https://doi.org/10.1109/TMC.2021.3120475>.
14. Xu, S.; Fan, K.K. A Silent Revolution: From Sketching to Coding—A Case Study on Code-Based Design Tool Learning. *EURASIA J. Math. Sci. Technol. Educ.* **2017**, *13*, 2959–2977. <https://doi.org/10.12973/eurasia.2017.00730a>.
15. Chen, K.C.; Lee, C.F.; Chang, T.W.; Wang, C.G.; Li, J.R. From Viewing to Structure: A Computational Framework for Modeling and Visualizing Visual Exploration. *Appl. Sci.* **2025**, *15*, 7900. <https://doi.org/10.3390/app15147900>.
16. Cao, J.; Zhao, W.; Hu, H.; Liu, Y.; Guo, X. Using Linkography and Situated FBS Co-Design Model to Explore User Participatory Conceptual Design Process. *Processes* **2022**, *10*, 713–732. <https://doi.org/10.3390/pr10040713>.
17. Siirtola, P.; Röning, J. Context-Aware Incremental Learning-Based Method for Personalized Human Activity Recognition. *J. Ambient Intell. Humaniz. Comput.* **2021**, *12*, 10499–10513. <https://doi.org/10.1007/s12652-020-02808-z>.
18. Theodoridou, E.; Cinque, L.; Mignosi, F.; Placidi, G.; Polsinelli, M.; Tavares, J.M.R.S.; Spezialetti, M. Hand Tracking and Gesture Recognition by Multiple Contactless Sensors: A Survey. *IEEE Trans. Hum. Mach. Syst.* **2023**, *53*, 35–43. <https://doi.org/10.1109/THMS.2022.3188840>.
19. Öztürk Kösençig, K.; Özbayraktar, M. Unveiling Interactions among Architectural Sketching, Parametric Design, and Digital Fabrication Using Linkography. *Int. J. Des. Creat. Innov.* **2025**, *13*, 1–22. <https://doi.org/10.1080/21650349.2024.2426646>.
20. Freitas, A.; Santos, D.; Lima, R.; Santos, C.G.; Meiguins, B. Pactolo Bar: An Approach to Mitigate the Midas Touch Problem in Non-Conventional Interaction. *Sensors* **2023**, *23*, 2110. <https://doi.org/10.3390/s23042110>.
21. Zeng, L.; Weber, G., User Interfaces for Pin-Array Tactile Displays. In *Advancements in Pin-Array Tactile Displays*; Springer Nature: Cham, Switzerland, 2025; pp. 29–45. https://doi.org/10.1007/978-3-032-03151-8_3.
22. Mummadi, C.; Leo, F.; Verma, K.; Kasireddy, S.; Scholl, P.; Kempfle, J.; Laerhoven, K. Real-Time and Embedded Detection of Hand Gestures with an IMU-based Glove. *Informatics* **2018**, *5*, 28–46. <https://doi.org/10.3390/informatics5020028>.
23. Dong, D.; Zhu, N.; Wang, J.; Li, Y. Lattice-Based Sensor Data Acquisition Strategy to Solve Sensor Position Drift in Human Gait Phase Recognition System with a Single Inertia Measurement Unit. *Eng. Appl. Artif. Intell.* **2025**, *147*, 110286. <https://doi.org/10.1016/j.engappai.2025.110286>.
24. Colli Alfaro, J.G.; Trejos, A.L. User-Independent Hand Gesture Recognition Classification Models Using Sensor Fusion. *Sensors* **2022**, *22*, 1321. <https://doi.org/10.3390/s22041321>.
25. Li, C.; Lee, C.F.; Xu, S. Stigma Threat in Design for Older Adults: Exploring Design Factors That Induce Stigma Perception. *Int. J. Des.* **2020**, Vol 14, No 1.
26. Kil, Y.S.; Lee, Y.J.; Jeon, S.E.; Oh, Y.S.; Lee, I.G. Optimization of Privacy-Utility Trade-off for Efficient Feature Selection of Secure Internet of Things. *IEEE Access* **2024**, *12*, 142582–142591. <https://doi.org/10.1109/ACCESS.2024.3467049>.
27. Lee, S.; Lim, Y.; Lim, K. Multimodal Sensor Fusion Models for Real-Time Exercise Repetition Counting with IMU Sensors and Respiration Data. *Inf. Fusion* **2024**, *104*, 102153. <https://doi.org/10.1016/j.inffus.2023.102153>.
28. Rivadeneira, J.E.; Borges, G.A.; Rodrigues, A.; Boavida, F.; Sá Silva, J. A Unified Privacy Preserving Model with AI at the Edge for Human-in-the-Loop Cyber-Physical Systems. *Internet Things* **2024**, *25*, 101034. <https://doi.org/10.1016/j.iot.2023.101034>.

29. Yemata, S.; Lemma, D.; Moussa, S.M. Exploring Factors and Features Impacting Data Privacy and Utility Trade-Offs in IoT-based Healthcare Systems: A Systematic Literature Review. *Inf. Commun. Soc.* **2025**, *28*, 3145–3174. <https://doi.org/10.1080/1369118X.2025.2492573>.
30. Lamaakal, I.; Essahraoui, S.; Maleh, Y.; Makkaoui, K.E.; Ouahbi, I.; Bouami, M.F.; El-Latif, A.A.A.; Almousa, M.; Peng, J.; Niyato, D. A Comprehensive Survey on Tiny Machine Learning for Human Behavior Analysis. *IEEE Internet Things J.* **2025**, *12*, 32419–32443. <https://doi.org/10.1109/JIOT.2025.3565688>.
31. Daniel, N.; Klein, I. INIM: Inertial Images Construction with Applications to Activity Recognition. *Sensors* **2021**, *21*, 4787. <https://doi.org/10.3390/s21144787>.
32. Wang, H.; Qiu, C.; Zhang, C.; Xu, J.; Su, C. P-CA: Privacy-preserving Convolutional Autoencoder-Based Edge–Cloud Collaborative Computing for Human Behavior Recognition. *Mathematics* **2024**, *12*, 2587. <https://doi.org/10.3390/math12162587>.
33. Kim, J.; Cho, S.H. A Differential Privacy Framework with Adjustable Efficiency–Utility Trade-Offs for Data Collection. *Mathematics* **2025**, *13*, 812–833. <https://doi.org/10.3390/math13050812>.
34. Zhou, H.; Zhang, X.; Feng, Y.; Zhang, T.; Xiong, L. Efficient Human Activity Recognition on Edge Devices Using DeepConv LSTM Architectures. *Sci. Rep.* **2025**, *15*, 13830. <https://doi.org/10.1038/s41598-025-98571-2>.
35. Li, J.; Liu, X.; Wang, Z.; Zhao, H.; Zhang, T.; Qiu, S.; Zhou, X.; Cai, H.; Ni, R.; Cangelosi, A. Real-Time Human Motion Capture Based on Wearable Inertial Sensor Networks. *IEEE Internet Things J.* **2022**, *9*, 8953–8966. <https://doi.org/10.1109/JIOT.2021.3119328>.
36. Ahmed, M.A.; Zaidan, B.B.; Zaidan, A.A.; Alamoodi, A.H.; Albahri, O.S.; Al-Qaysi, Z.T.; Albahri, A.S.; Salih, M.M. Real-Time Sign Language Framework Based on Wearable Device: Analysis of MSL, DataGlove, and Gesture Recognition. *Soft Comput.* **2021**, *25*, 11101–11122. <https://doi.org/10.1007/s00500-021-05855-6>.
37. Pan, T.Y.; Chang, C.Y.; Tsai, W.L.; Hu, M.C. Multisensor-Based 3D Gesture Recognition for a Decision-Making Training System. *IEEE Sens. J.* **2021**, *21*, 706–716. <https://doi.org/10.1109/JSEN.2020.3012887>.
38. Li, J.; Huang, L.; Shah, S.; Jones, S.J.; Jin, Y.; Wang, D.; Russell, A.; Choi, S.; Gao, Y.; Yuan, J.; et al. SignRing: Continuous American Sign Language Recognition Using IMU Rings and Virtual IMU Data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2023**, *7*, 1–29. <https://doi.org/10.1145/3610881>.
39. Qureshi, T.S.; Shahid, M.H.; Farhan, A.A.; Alamri, S. A Systematic Literature Review on Human Activity Recognition Using Smart Devices: Advances, Challenges, and Future Directions. *Artif. Intell. Rev.* **2025**, *58*, 276. <https://doi.org/10.1007/s10462-025-11275-x>.
40. He, Z.; Sun, Y.; Zhang, Z. Human Activity Recognition Based on Deep Learning Regardless of Sensor Orientation. *Appl. Sci.* **2024**, *14*, 3637. <https://doi.org/10.3390/app14093637>.
41. Shavit, Y.; Klein, I. Boosting Inertial-Based Human Activity Recognition with Transformers. *IEEE Access* **2021**, *9*, 53540–53547. <https://doi.org/10.1109/ACCESS.2021.3070646>.
42. Hoang, M.L. A Review of Developments and Metrology in Machine Learning and Deep Learning for Wearable IoT Devices. *IEEE Access* **2025**, *13*, 106035–106054. <https://doi.org/10.1109/ACCESS.2025.3573937>.
43. Müller, P.N.; Müller, A.J.; Achenbach, P.; Göbel, S. IMU-based Fitness Activity Recognition Using CNNs for Time Series Classification. *Sensors* **2024**, *24*, 742. <https://doi.org/10.3390/s24030742>.
44. Hashi, A.O.; Hashim, S.Z.M.; Asamah, A.B. A Systematic Review of Hand Gesture Recognition: An Update from 2018 to 2024. *IEEE Access* **2024**, *12*, 143599–143626. <https://doi.org/10.1109/ACCESS.2024.3421992>.
45. Just, F.; Ghinami, C.; Zbinden, J.; Ortiz-Catalan, M. Deployment of Machine Learning Algorithms on Resource-Constrained Hardware Platforms for Prosthetics. *IEEE Access* **2024**, *12*, 40439–40449. <https://doi.org/10.1109/ACCESS.2024.3371251>.
46. Liu, H.I.; Galindo, M.; Xie, H.; Wong, L.K.; Shuai, H.H.; Li, Y.H.; Cheng, W.H. Lightweight Deep Learning for Resource-Constrained Environments: A Survey. *ACM Comput. Surv.* **2024**, *56*, 267. <https://doi.org/10.1145/3657282>.
47. Rahman, S.; Pal, S.; Yearwood, J.; Karmakar, C. Analysing Performances of DL-based ECG Noise Classification Models Deployed in Memory-Constraint IoT-enabled Devices. *IEEE Trans. Consum. Electron.* **2024**, *70*, 704–714. <https://doi.org/10.1109/TCE.2024.3370709>.
48. Gill, S.S.; Golec, M.; Hu, J.; Xu, M.; Du, J.; Wu, H.; Walia, G.K.; Murugesan, S.S.; Ali, B.; Kumar, M.; et al. Edge AI: A Taxonomy, Systematic Review and Future Directions. *Clust. Comput.* **2025**, *28*, 18–61. <https://doi.org/10.1007/s10586-024-04686-y>.
49. Pawłowski, M.; Wróblewska, A.; Sysko-Romańczuk, S. Effective Techniques for Multimodal Data Fusion: A Comparative Analysis. *Sensors* **2023**, *23*, 2381. <https://doi.org/10.3390/s23052381>.
50. Malinvern, L.; Schaper, M.M.; Pares, N. Multimodal Methodological Approach for Participatory Design of Full-Body Interaction Learning Environments. *Qual. Res.* **2019**, *19*, 71–89. <https://doi.org/10.1177/1468794118773299>.
51. Badidi, E.; Moumane, K.; Ghazi, F.E. Opportunities, Applications, and Challenges of Edge-AI Enabled Video Analytics in Smart Cities: A Systematic Review. *IEEE Access* **2023**, *11*, 80543–80572. <https://doi.org/10.1109/ACCESS.2023.3300658>.
52. Xaviar, S.; Yang, X.; Ardakanian, O. Centaur: Robust Multimodal Fusion for Human Activity Recognition. *IEEE Sens. J.* **2024**, *24*, 18578–18591. <https://doi.org/10.1109/JSEN.2024.3388893>.

53. Tzimos, N.; Parafestas, E.; Voutsakelis, G.; Kontogiannis, S.; Kokkonis, G. Multimodal Interaction with Haptic Interfaces on 3D Objects in Virtual Reality. *Electronics* **2025**, *14*, 4035. <https://doi.org/10.3390/electronics14204035>.
54. Chang, T.W. Supporting Design Learning with Design Puzzles: Some Observations of on-Line Learning with Design Puzzles. In *Recent Advances in Design and Decision Support Systems in Architecture and Urban Planning*; Van Leeuwen, J.P., Timmermans, H.J.P., Eds.; Springer: Dordrecht, The Netherlands, 2004; pp. 293–307. https://doi.org/10.1007/1-4020-2409-6_19.

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.