

Article

PMCI: A Prototype-Based Diagnostic Index for Cross-Modal Affective Agreement

Yernar Seksenbayev ^{1, }, Saule Kudubayeva ^{1 }, Abdykarim Baimankulov ^{2 }, Aigerim Yerimbetova ^{3,4,* },
Elmira Daiyrbayeva ^{3,5,* }, Ulmeken Berzhanova ^{3,6 } and Bakzhan Sakenov ^{3 }

- ¹ Faculty of Digital Sciences and Artificial Intelligence, L.N. Gumilyov Eurasian National University, Astana 010008, Kazakhstan; simpleuser6501@gmail.com (Y.S.); saule.kudubayeva@gmail.com (S.K.)
² Department of Information Systems, A. Baitursynov Kostanay Regional University, Kostanay 110000, Kazakhstan; bat_56@mail.ru
³ Institute of Information and Computational Technologies CS MSHE RK, Almaty 050000, Kazakhstan; berzhanovaulmekenn@gmail.com (U.B.); sbakzhan22@gmail.com (B.S.)
⁴ School of Engineering and Information Technology, META University, Almaty 050012, Kazakhstan
⁵ Department of Software Engineering, Satbayev University, Almaty 050010, Kazakhstan
⁶ Faculty of Information Technology and Artificial Intelligence, Farabi University, Almaty 050038, Kazakhstan
* Correspondence: aigerian8888@gmail.com (A.Y.); e.daiyrbayeva@satbayev.university (E.D.)

Abstract

We introduce the Probabilistic Multimodal Consistency Index (PMCI), a prototype measure of probabilistic semantic agreement used to quantify how much two modalities share in terms of mutually compatible affective evidence. PMCI does not model feature fusion or emotion classification directly; instead, each modality is mapped to a different probability distribution over learnable latent affective-agreement anchors, and the agreement of those distributions is quantified via the Jensen–Shannon divergence. Consequently, we propose PMCI not as a substitute for a discriminative model of pair matching, but rather as an auxiliary diagnostic index of cross-modal affective agreement. Experiments were conducted on pose-based facial and body keypoint sequences obtained from the RAVDESS and MELD datasets. In the updated RAVDESS setup, we extended the cache of preprocessed keypoint files to include all 24 actors, and we sampled four distinct pose-derived 12-frame crops per source video. This resulted in a total of 11,520 pose-derived windows, with the actor label determining the train/validation/test splits for the RAVDESS dataset. MELD contained 402 filtered pose-derived windows and served as an auxiliary in-the-wild validation setting with additional noise. When applying the updated RAVDESS standard pair-matching, DirectCosine_K0 achieved ROC–AUC = 0.943 and PR–AUC = 0.917, demonstrating that it is indeed the best exact pair-matching baseline and that PMCI-based configurations are not as accurate as DirectCosine_K0 when performing common exact face–body pair discrimination. Under this protocol, PMCI_K8, PMCI_K16, and PMCI_K32 produced ROC–AUC scores of 0.851, 0.805, and 0.887, respectively. Running a one-window-per-source-video experiment with 100 repetitions yielded similar results in the following order: DirectCosine_K0 = 0.942 plus-minus 0.009, PMCI_K16 = 0.807 plus-minus 0.016, and PMCI_K32 = 0.893 plus-minus 0.012, which shows that the results obtained were not the result of only repeated temporal crops. In the DirectCosine-mined hard negative setting, PMCI produced modest diagnostic separation above chance, where Direct+PMCI_K32 yielded ROC–AUC = 0.582. In a suite of shuffle control, permutation control, model initialization control, temperature control, anchor usage control, control latency, and pose perturbation control experiments, the diagnostic stability of PMCI was tested, confirming that PMCI is sensitive to both pose-estimation accuracy and domain shift. Finally, PMCI must be interpreted as a probabilistic diagnostic index of cross-modal affective agreement that is based on pose, not as a general-purpose emotion-recognition system.



Academic Editor: Firstname Lastname

Received: 25 June 2026

Revised: 16 July 2026

Accepted: 17 July 2026

Published: 19 July 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.

Keywords: multimodal affective computing; probabilistic consistency; facial and gestural cues; cross-modal agreement; interactive systems; affective coherence; probabilistic anchor distributions; Jensen–Shannon divergence; pose-based analysis; hard-negative evaluation

1. Introduction

Most prior work in multimodal affective computing has focused on recognition performance, fusion strategies, or modality alignment for downstream prediction. However, in interactive systems, it is also important to determine whether different expressive channels convey mutually consistent affective interpretations. This distinction motivates a separate line of inquiry: multimodal consistency analysis. From this perspective, facial and gestural signals are not only complementary inputs for classification, but also partially independent affective observations whose agreement or disagreement may itself carry useful information. The present study builds on this view by positioning probabilistic cross-modal consistency as an independent object of analysis rather than as a secondary outcome of multimodal fusion. In interactive systems, a user's affective state is expressed through several channels at the same time, including facial expression, body posture, gesture, vocal tone, and situational context. For this reason, combining modalities only to improve recognition accuracy is not sufficient. It is also necessary to determine whether these modalities support the same affective interpretation or whether they contain a meaningful mismatch.

The present study does not tackle the issue of multimodal data fusion, although it was inspired by research in the area of multimodal affective computing. Instead, PMCI addresses a distinct problem, which is how to quantify a degree of agreement between two modality-specific interpretations represented in the form of a probability distribution. Accordingly, PMCI is not a method for feature fusion but, rather, a probabilistic semantic agreement metric for comparing facial and gestural probability distributions inferred independently from different sources of evidence.

Despite growing interest in multimodal affective computing, an important gap remains between multimodal fusion and multimodal consistency analysis. Fusion-oriented approaches are primarily designed to improve downstream recognition performance by combining heterogeneous sources of information, but they do not necessarily provide an explicit account of whether different modalities support the same affective interpretation. In interactive systems, this distinction is important: a model may produce a confident prediction while the underlying channels remain weakly aligned or even partially contradictory. This is particularly relevant in settings involving subtle expression, emotional masking, hesitation, or controlled social display.

The contribution of this work is the formulation of cross-modal affective agreement as a distribution-level comparison problem. Facial and gestural streams are not fused into a single representation before comparison. Instead, each modality produces its own probability distribution over a shared latent anchor space, and PMCI quantifies the agreement between these two distributions. This formulation separates semantic agreement estimation from ordinary exact pair matching, for which direct embedding similarity or discriminative fusion may remain more effective.

The central hypothesis is that agreement between modality-specific probability distributions can provide a useful diagnostic coordinate in cases where direct embedding similarity becomes ambiguous. Direct similarity can be effective for identifying whether two streams originate from the same video segment, but it may also reflect actor identity, motion style, or other shortcut-related correspondences. PMCI is therefore evaluated as a continuous semantic-agreement score rather than as a discriminative pair-matching classifier.

The results of this study can be characterised by five key observations. Firstly, both DirectCosine and LateFusionMLP continue to serve as robust baselines for ordinary exact face–body pair matching. Secondly, PMCI should be regarded not as a substitute classifier, but rather as a probabilistic semantic-agreement measure over shared latent anchors. Thirdly, PMCI is most relevant in hard-negative scenarios where high embedding similarity is no longer reliable.

The experiments are organized to separate ordinary exact pair matching from cross-modal semantic agreement. DirectCosine and LateFusionMLP are retained as discriminative baselines for exact face–body matching. PMCI is evaluated separately as a diagnostic agreement coordinate, with particular attention to DirectCosine-mined hard negatives, shuffled and random controls, rank-balanced hard-negative evaluation, and decile-based low-agreement analysis.

In uncertainty-aware interaction monitoring, PMCI may be used as an auxiliary diagnostic signal rather than as a final emotion label. A low PMCI score indicates that the facial and gestural streams produce weakly compatible probability distributions. Such cases may be routed to a more cautious decision policy, additional contextual analysis, or human review. This interpretation is limited to the pose-based facial and body representations evaluated in the present study.

The remainder of this paper is organized as follows. Section 2 reviews related work on face–body affect perception, multimodal emotion recognition, semantic similarity, and probability-distribution agreement. Section 3 presents the PMCI formulation, task definition, datasets, and experimental protocol. Section 4 reports ordinary matching results, hard-negative diagnostics, non-redundancy controls, rank-balanced evaluation, decile analysis, and ablation studies. Section 5 discusses methodological implications and limitations. Section 6 concludes the paper.

2. Related Works

Classical theories of emotion, including those associated with Darwin and Ekman, primarily emphasized prototypical facial expressions of several basic emotions [1,2]. However, bodily signals also play a significant role in affective communication. It has been demonstrated that they can enhance the transmission of facial expression, thereby facilitating enhanced interpretation of their intended meaning or modifying the comprehensive analysis of the observed configuration. From this perspective, the face and body should not be considered as isolated feature sources; instead, they should be understood as interconnected expressive channels that jointly contribute to affective meaning. Plutchik’s model posits the hypothesis that emotions can be conceptualized as parameters of a structured space of categories and intensities, which can manifest through a series of mutually expressive channels [3]. Dimensional approaches provide a description of affect through valence and arousal, whereas appraisal-based frameworks predominantly connect emotion as an evaluation of the situation and context [4,5]. The aforementioned approaches underscore the notion that the affective state should not be reduced to a solitary observable signal, and that signals can express representation through multiple modalities, with each modality providing a partial interpretation of a more multifaceted affective state. For instance, while downcast and bowed postures could be a non-verbal indicator of sadness, expressive gestures and high posture could serve as non-verbal indicators of joy. Thus, bodily expressions could act to disambiguate the facial expression, thereby allowing for a more robust estimation of the overall emotion. This finding indicates that, in contrast to the independent modelling of facial features and gesture in conventional approaches, interactive systems should adopt a more holistic approach by jointly modelling both features and gesture, with a view to enhancing mutual compatibility in the interpretation of emotion.

Neurophysiological data additionally confirm early integration between facial and bodily signals. Meeren et al. showed that visual processing in the brain had modulation approximately during 115 ms depending on whether the face and body were emotionally consistent [6]. This early effect suggests that consistency between the face and the body is a significant component of emotion perception and interpretation. Accordingly, analysis of the consistency between signals of the face and gestures can be understood as a computational formulation that is consistent with canonical mechanisms of multimodal perception of affect.

Beyond controlled experiments, real-life emotional behavior often involves nuanced coordination (or lack thereof) between modalities. Body movements can effectively reflect changes in affective state even when facial cues are subtle. Observers tend to rely more on body expressions than on facial expressions or voice tone under high emotional intensity, ambiguity, or when cues conflict. Indeed, bodily expressions may dominate in such scenarios: Aviezer et al. found that body cues, not faces, distinguished positive vs. negative intense emotions [7]. The influence of the body on face-based emotion recognition is generally stronger [8]. These human perceptual tendencies underscore the importance of quantifying facial–gestural affective coherence in computational systems.

Multimodal emotion recognition has accordingly evolved from uni-modal analysis to more integrative approaches. Barros and Wermter developed a cross-channel model to integrate facial and auditory cues for expression recognition [9], and more recently Hu et al. proposed a prototype-grounded approach for multimodal emotion recognition, conceptually related to prototype-based alignment [10]. Despite these advances, current systems typically fuse or align modalities at the feature or embedding level, and often assume that the “true” emotion is consistently expressed across modalities. In reality, however, facial and bodily signals can be asynchronous or even divergent; emotion coherence across channels is not guaranteed and can be surprisingly low for certain emotions.

In human perception studies, this agreement has been described in terms of emotion coherence or congruency, and its absence is known to hinder emotion recognition. In character animation, ensuring that a virtual avatar’s facial expressions match its body language greatly enhances perceived realism. Chhatre et al. showed that users strongly prefer animations where facial and bodily channels are congruently driven compared to disjoint or single-modality animations [11]. In human–computer interaction and social robotics, a mismatch between a person’s face and gesture cues might indicate suppressed emotion, sarcasm, or deceptive behavior.

Previous perceptual studies show that facial and bodily expressions are jointly involved in emotion interpretation. Congruent face–body modality pairs were connected with more competent recognition and faster responses in certain experimental conditions [12]. The body can not only complement the face, but also change its emotional meaning. In the experiments by Aviezer et al., the same facial expression caused both positive and negative labels depending on the bodily context with which it was connected [7]. Lecker et al. also reported a bidirectional contextual effect, in which facial expression can influence the perceived bodily emotion; however, the influence of the body on the face often turns out stronger in certain situations [8].

These results show that emotional congruence is a relation between modalities, not a property of one separate modality. High congruence forms a more holistic perception, since facial and bodily signals support a similar affective interpretation. Inconsistency, on the contrary, can lead to mixed or unstable interpretation. In real behavior, complete consistency is not guaranteed. Emotional expression can be masked, regulated by social norms, suppressed, or manifested only partially. Reisenzein et al. showed that external facial expressions do not always reliably accompany experienced emotion [13]. Therefore,

the proposed approach does not assume that emotion signals are in every case congruent between modalities; instead, it considers congruence as a single measurable signal.

In parallel with this, affective computing developed from unimodal analysis to multimodal fusion. Poria et al. show that combining facial expressions, speech intonation, and text transcripts could increase the robustness of emotion recognition compared with separate modalities [14]. Deep learning allowed the development of fusion strategies through attention mechanisms and coordinated representations. Zadeh et al. proposed tensor fusion for tasks of multimodal sentiment analysis and emotion determination [15], whereas Tsai et al. used cross-modal attention for aligning latent dynamics in unaligned multimodal sequences [16]. These methods are important for classification and prediction; however, they do not directly define a distribution-level measure of semantic agreement between modality-specific interpretations. In such models, cross-modal consistency is usually treated only implicitly, insofar as it improves the performance of the final prediction task. This leaves open the separate problem of measuring semantic agreement between modality-specific interpretations themselves.

The proposed formulation is also related to research on semantic similarity and semantic equivalence. In this literature, the central problem is not whether two observations are identical, but whether they support related or compatible meanings. Previous studies have addressed this question in different settings, including fuzzy relational databases, document-level semantic similarity, and general surveys of semantic similarity approaches and metrics [17–19]. In contrast to document-level semantic similarity, the present study compares semantic interpretations inferred from non-verbal visual modalities. PMCI is also conceptually related to probability-distribution agreement and expert-consensus approaches [20]. In such settings, agreement can be estimated by comparing probability distributions through overlap, divergence, or aggregation-based measures. The present work follows a similar distribution-level logic, but the compared distributions are not expert probability judgments. They are modality-specific affective distributions inferred from facial and gestural pose evidence. Therefore, PMCI can be understood as a probabilistic semantic-agreement measure applied to cross-modal affective evidence rather than as a conventional feature-fusion method.

Recognition of emotions based on pose is the closest approach to the problem considered here, working with body movement and pose keypoints of the silhouette [19]. Movement of body pose is a significant nonverbal channel of affective communication, and previous reviews address both psychological and computational approaches to perception and recognition of bodily expression [21]. Kinematics and pose also play an important role in emotion perception, which in general justifies the use of temporal derivatives as support to pose coordinates [22]. Pose-estimation methods such as OpenPose further motivate the use of body keypoints for visual behavior analysis [23]. Neuroscientific approaches also show that bodily expressions carry affective meaning and can be effectively processed [24].

To address this issue, a two-stage probabilistic formulation is proposed. In the first stage, each modality estimates a probability distribution over a shared set of learnable prototype anchors. In the second stage, the consistency of the two modality-specific distributions is compared to evaluate compatibility across modalities.

Although there are specialized problems involving audiovisual temporal alignment and synchronization [25], these are different and distinct from the problem of short-window consistency examined here. Koelstra et al. introduced DEAP, a multimodal benchmark that illustrates practical issues in emotion analysis across heterogeneous signals [26]. More directly, the visual analytics system EmoCo extracted emotions independently from several modalities and examined their coherence over time [27], which conceptually motivates the present approach [28].

3. Methods

3.1. Datasets and Model Framework

We evaluated the proposed framework on two external video datasets: RAVDESS [29] and MELD [30]. RAVDESS was used as a controlled acted-emotion dataset, allowing subject-independent actor-level evaluation and emotion-controlled counterfactual analysis. MELD was used as a noisy in-the-wild conversational dataset to test whether the proposed consistency score remains informative under less controlled visual conditions. For the revised experiments, the RAVDESS preprocessing cache was expanded to all 24 actors. Instead of extracting a single 12-frame sequence from each source video, four temporal 12-frame crops were extracted per source video. This produced 11,520 pose-derived RAVDESS windows, split by actor into 8160 training windows, 1920 validation windows, and 1440 test windows. MELD retained 402 pose-derived windows after filtering and was split by video into training, validation, and test partitions. Because multiple temporal crops can originate from the same source video, these windows are not treated as fully independent behavioral samples. For this reason, additional source-video-aware controls were included, including a one-window-per-source-video control and source-video/actor grouped bootstrap analyses.

The proposed approach is organized as a pose-based multimodal pipeline. The overall workflow of the proposed PMCI framework is shown in Figure 1. Initially, facial and body keypoints are extracted from video frames using YOLOv8, thereby enabling the evaluation to be conducted more objectively. Subsequently, the obtained keypoint sequences are subjected to normalisation and augmented with short-term temporal differences.

Facial and body modalities are encoded by separate GRU branches, after which their latent representations are projected into a shared prototype-anchor space. The modalities generate their own probability distributions across the shared learnable anchors, and the resulting PMCI score is produced by quantifying agreement between the distributions via Jensen–Shannon divergence. There are four distinct processing stages in the architecture: pose extraction; temporal encoding; probabilistic prototype-anchor representation; and estimation of cross-modal semantic agreement.

For each frame, we obtained pose keypoint features using an off-the-shelf YOLOv8 pose estimator [31]. We extracted face and body keypoints, represented as normalized coordinates. The face pose at time t is represented as a 10-dimensional vector $\mathbf{x}_t^F$ and the body pose as a 24-dimensional vector $\mathbf{x}_t^B$. These coordinate features are enriched with temporal derivatives. In our final setup, we included two past frames' differences for each keypoint coordinate. The augmented representations are formed as:

$$\mathbf{f}_t = [\mathbf{x}_t^F, \mathbf{x}_t^F - \mathbf{x}_{t-1}^F, \mathbf{x}_t^F - \mathbf{x}_{t-2}^F], \quad \mathbf{b}_t = [\mathbf{x}_t^B, \mathbf{x}_t^B - \mathbf{x}_{t-1}^B, \mathbf{x}_t^B - \mathbf{x}_{t-2}^B], \quad (1)$$

yielding a 30-D face feature vector and a 72-D body feature vector per frame. All features are standardized across the dataset.

As the design indicates, our method is explicitly pose-based: our feature vector is defined using facial and body keypoint geometry and the short-term temporal changes in their coordinates, but it lacks facial RGB texture, speech, vocal prosody, and text. Therefore, the PMCI approach implemented in this paper should be regarded as a pose-derived visual agreement model, and not as a fully multimodal affect recognition system. The use of short-term temporal changes captures only the local motion within each analysis window, and it is not meant to model either long-range affective dynamics or longer temporal misalignments.

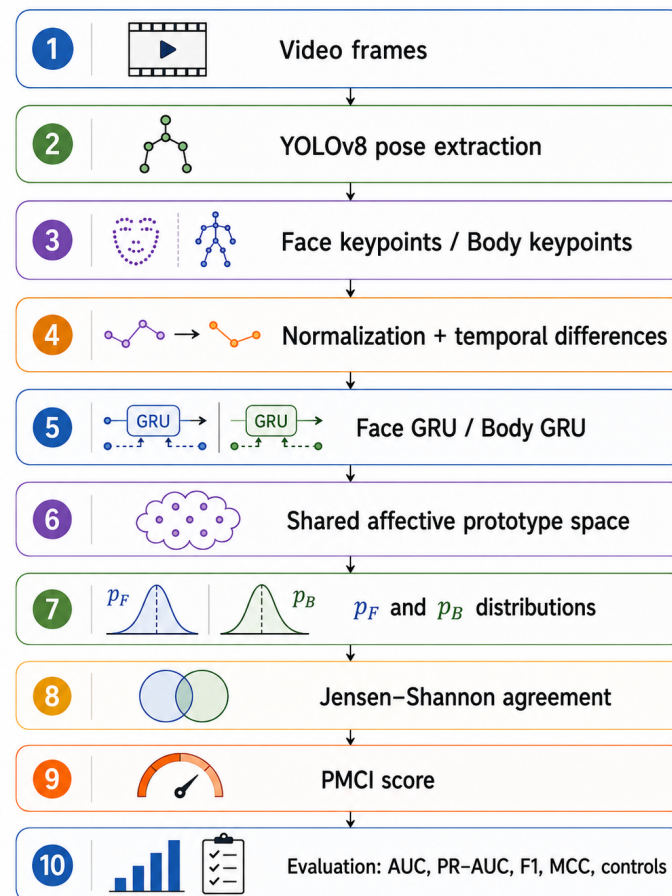


Figure 1. General workflow of the proposed PMCI framework.

The use of the approach of collecting features obtained from pose, instead of using raw RGB frames, is motivated by the necessity of more stable experimental control and clearer interpretation of results. The use of direct raw RGB video, in which the model can reach a high matching score by relying on such usual features as actor appearance, clothing color, background, lighting, image texture or other shortcut-features of background level. Such information can increase the effectiveness of cross-modal matching, but this does not necessarily reflect affective consistency between face and body.

For this reason, in the present study a pose-based representation is used, it limits the model to features more closely connected with nonverbal affective expression. Pose-based representation focuses the analysis on the location of keypoints in the frame, body geometry, pose and movement. These features are usually more suitable for analysis of facial-gestural consistency, since affective expression manifests through configuration and dynamics.

Temporal derivatives allow the model to represent not only static positions of keypoints in space, but also short-term expressive dynamics, including direction of movement, onset of movement, and local intensity of changes. In this sense, pose representation performs the function of a controlled intermediate layer between raw observation of emotions and probabilistic affective interpretation.

The use of keypoints also reduces the influence of color, texture, identity, and background cues. This is especially important for the present task because the model should not match facial and gestural streams through actor identity or visual properties of the scene. It should evaluate the extent to which the two modalities are consistent as expressive channels. This design does not remove all possible shortcuts, but it substantially reduces the risk that a high score is obtained from irrelevant visual correspondence.

To model multimodal affective consistency, the proposed Probabilistic Multimodal Consistency Network uses a dual-branch architecture. The facial branch encodes facial feature dynamics, while the gestural branch encodes bodily feature dynamics. Both branches are then projected into a shared affective prototype space. Unlike direct latent matching, the model does not force facial and gestural embeddings to coincide directly. Instead, each modality produces a modality-specific probability distribution over shared affective prototypes.

The temporal encoders in this architecture are standard sequence-modeling components. The Facial Encoder processes a sequence of 30-D face features $\{\mathbf{f}_t\}_{t=1}^T$ using a GRU [32]. The Gestural Encoder processes a sequence of 72-D body/gesture features $\{\mathbf{b}_t\}_{t=1}^T$ using a parallel GRU [32].

Our two branches output modality-specific latent representations, which are matched using a common set of K learnable latent affective-agreement anchors. The resulting similarity to each anchor is passed through a temperature-scaled softmax to generate modality-specific probability distributions p^F and p^B . The hyperparameter K defines the granularity of our shared prototype-anchor space; it is not intended to be regarded as the number of distinct emotion categories in our emotion ontology, and each prototype anchor is not guaranteed to have a one-to-one correspondence with an emotion in human psychology. Rather, these prototype anchors are learned operational reference points that we use to form our distributions for the purpose of agreement estimation. In our main experiments, we tested $K \in \{8, 16, 32\}$, while the $K = 0$ condition omits the anchor layer as a baseline to measure direct embedding similarity rather than representing a PMCI model. We also tested $K \in \{4, 8, 16, 32, 64\}$ in stability checks to determine whether the anchor space is too coarse- or fine-grained to produce stable agreement scores.

The methodological novelty lies not in the GRU itself, but in the formulation of cross-modal comparison. In conventional matching approaches, two modalities are compared by direct geometric proximity between embeddings. In the proposed model, facial and gestural streams are first converted into probabilistic interpretations through a shared prototype space. Consistency is then defined through Jensen–Shannon agreement between p_F and p_B .

The two GRU branches produce latent affective representations:

$$\mathbf{z}_F = \text{GRU}_F(\mathbf{f}_{1:T}), \quad \mathbf{z}_B = \text{GRU}_B(\mathbf{b}_{1:T}). \quad (2)$$

These latent representations are compared against the shared set of K learnable latent affective anchors $\{c_k\}_{k=1}^K$. For each modality, anchor similarities are converted into probability distributions via temperature-scaled softmax:

$$p_F^{(k)} = \frac{\exp(\text{sim}(\mathbf{z}_F, c_k)/\tau)}{\sum_{j=1}^K \exp(\text{sim}(\mathbf{z}_F, c_j)/\tau)}, \quad p_B^{(k)} = \frac{\exp(\text{sim}(\mathbf{z}_B, c_k)/\tau)}{\sum_{j=1}^K \exp(\text{sim}(\mathbf{z}_B, c_j)/\tau)}, \quad (3)$$

where $\tau > 0$ is a temperature parameter controlling the sharpness of the prototype distributions, and $\text{sim}(\cdot, \cdot)$ denotes cosine similarity. Within the proposed probabilistic multimodal consistency framework, the Probabilistic Multimodal Consistency Index (PMCI) is defined as:

$$\text{PMCI}(\mathbf{z}_F, \mathbf{z}_B) = 1 - \text{JSD}(p_F \parallel p_B), \quad (4)$$

where $\text{JSD}(\cdot \parallel \cdot)$ denotes the Jensen–Shannon divergence between the modality-specific probability distributions inferred from the facial and gestural streams, respectively [33].

The Jensen–Shannon divergence is chosen because it is symmetric, bounded, and defined for discrete probability distributions over a common anchor space [33]. Therefore, it is suitable for comparing the complete modality-specific probability distributions, rather than simply the distance between two points. The selection of the Jensen–Shannon divergence does not assume its uniqueness, and so we conducted further robustness checks by evaluating other divergence and overlap measures including Hellinger distance, total variation distance, probability dot product, and probability cosine. We want our prototype-anchor space to be regarded as a latent operational agreement space that is not guaranteed to have any preexisting semantically aligned emotion ontology. Our facial and gesture GRU branches are trained to the same set of K learnable latent anchors, so we expect their probability distributions to become more comparable using the pair consistency loss, but this does not guarantee in a mathematical sense that each anchor in each modality has identical human-interpretable semantic meaning. We therefore validate this empirically using our matched/mismatched pair evaluation, disruption control experiments, shuffle control experiments, hard-negative test experiments, and prototype-anchor usage metrics. Note that our model does not treat examples where p_F and p_B assign all of their mass to disjoint anchors as errors or failures, but rather they indicate a low PMCI score corresponding to a low degree of cross modal agreement.

The potential application of PMCI lies in uncertainty-aware human–computer interaction. A classifier such as DirectCosine or LateFusionMLP can support exact pair matching or emotion classification, while PMCI functions as an auxiliary diagnostic metric that evaluates the congruence of emotional expressions conveyed through facial and bodily gestures. A low PMCI score may signify ambiguity, masking, transience, or weak alignment. An interactive system can use this information to delay an affective response, request human intervention, or indicate uncertainty rather than forcing a single emotion label.

The complete training and evaluation procedure is summarized in Algorithm 1.

Algorithm 1 Training and evaluation procedure of the proposed PMCI framework.

Require: Facial sequence F , body sequence B , number of anchors K , temperature parameter τ , positive and negative facial–gestural pairs

Ensure: Trained facial encoder, gestural encoder, latent affective-agreement anchors, and PMCI scores

- 1: Extract normalized facial and body pose keypoints from each video frame.
 - 2: Construct temporal feature sequences using coordinate features and short-term temporal differences.
 - 3: Form genuine facial–body pairs and mismatched negative pairs.
 - 4: Initialize the facial GRU encoder, gestural GRU encoder, and K learnable latent affective-agreement anchors.
 - 5: **for** each training epoch **do**
 - 6: **for** each mini-batch **do**
 - 7: Encode the facial and body sequences using separate GRU branches.
 - 8: Compare the resulting latent vectors with the shared anchor space.
 - 9: Convert prototype similarities into modality-specific probability distributions.
 - 10: Compute the PMCI score using Jensen–Shannon agreement.
 - 11: Update model parameters by increasing PMCI for genuine pairs and decreasing it for mismatched pairs.
 - 12: **end for**
 - 13: **end for**
 - 14: Evaluate the trained model using ROC–AUC, PR–AUC, F1-score, MCC, and related binary metrics.
-

The model was trained using a binary pair-consistency objective. Training used binary cross-entropy with logits over a trainable affine transformation of the PMCI score for

prototype-based variants, or of cosine similarity for the $K = 0$ baseline. Four negative pairing schemes were implemented: cross-video mismatches, same-video hard negatives, temporal-shift mismatches, and random shuffling. Optimization was performed using AdamW [34].

The present evaluation does not treat temporal alignment as a separate sequence-registration problem. Temporal-shift experiments are used as diagnostic perturbation controls rather than as evidence that PMCI solves long-range temporal misalignment. Therefore, the reported results should be interpreted as short-window facial–gestural agreement estimates rather than as full temporal-dynamics modeling.

In the revised evaluation, performance was assessed using the following metrics: ROC–AUC [35], PR–AUC, accuracy, balanced accuracy, precision, recall, F1-score, Matthews correlation coefficient (MCC), and specificity. The computation of threshold-dependent metrics was conducted on the basis of thresholds selected on the validation split, which were subsequently applied to the test split.

3.2. Experimental Protocol and Controls

The evaluation was designed to separate exact pair discrimination from probabilistic semantic consistency. First, subject-independent validation was performed on RAVDESS using actor-level splits. Second, the method was evaluated on MELD as a noisy in-the-wild condition. Third, cross-dataset transfer from RAVDESS to MELD was tested, together with a target-domain adaptation curve using 0%, 10%, 20%, 50%, and 100% of the MELD training subset.

To verify the presence of shortcut learning, we incorporated face-zero, body-zero, and body-shuffle control experiments, along with a linear actor-identity classifier to assess identity leakage. We also conducted balanced counterfactual experiments across true pairs, same-emotion/different-actor pairs, same-actor/different-emotion pairs, different-emotion/different-actor pairs, and random mismatches. All experiments were executed using three random seeds, and results are reported as mean $\pm$ standard deviation. The experimental design, settings, and hyperparameters are summarised in Table 1 and the configuration block that follows it.

Table 1. PMCI experimental design and evaluation protocol.

Evaluation Layer	Purpose	Experimental Setting
Input and datasets	Define the pose-based evidence used in the experiments	RAVDESS: 11,520 pose-derived 12-frame windows from all 24 actors, using four temporal crops per source video; split by actor into 8160 train, 1920 validation, and 1440 test windows. MELD: 402 pose-derived windows after filtering; split by video.
Core model comparison	Separate exact pair matching from probabilistic agreement estimation	DirectCosine_K0, LateFusionMLP, PMCI_K8, PMCI_K16, PMCI_K32, and Direct+PMCI variants.
Anchor-space granularity	Test the effect of K on PMCI behavior	K controls the granularity of the latent affective-agreement anchor space and is not interpreted as the number of psychological emotion categories.
Pair-bank and hard-negative evaluation	Test PMCI under balanced and difficult mismatch conditions	Balanced pair-bank evaluation; DirectCosine-mined hard negatives; one-window-per-source-video control; source-video and actor grouped bootstrap. Shuffled-PMCI controls, random/untrained controls, modality disruption, temporal perturbation, identity-leakage analysis, identity-excluded hard negatives, K-ablation, tau sensitivity, initialization checks, latency, and pose-perturbation checks.
Controls and robustness checks	Test whether PMCI provides non-redundant information	

3.3. Reviewer-Oriented Stress Tests and Non-Redundancy Controls

In addition to standard evaluation metrics, we evaluated whether PMCI offers non-redundant information beyond DirectCosine. A balanced test set of 48,000 pairs contained 24,000 genuine and 24,000 mismatched pairs. A separate hard-negative set contained 240 genuine pairs and 4800 hard negatives selected as the 20 most similar incorrect streams under DirectCosine for each anchor. A multi-shuffle control with 100 shuffled-PMCI variants tested whether any improvement stemmed from real cross-modal structure rather than additional feature dimensions. Fair hard-negative mining was also carried out using DirectCosine, PMCI_K8, and PMCI_K16 as mining targets. OOD perturbation tests included complete modality removal, time reversal, time shift, and partial corruption.

The expanded RAVDESS cache increases the number of pose-derived windows and includes all 24 actors, but multiple temporal crops from the same source video should not be treated as fully independent behavioral samples. To address this issue, we added a one-window-per-source-video control, in which only one randomly selected crop from each source video was used during evaluation. We also report source-video and actor grouped bootstrap analyses for hard-negative scores. These controls do not replace validation on larger naturalistic datasets, but they reduce the risk that the reported effects are explained only by repeated crops from the same videos. Overfitting risk was further limited through actor-independent RAVDESS splits, video-level MELD splits, validation-based threshold selection, pose-only features, and reviewer-oriented stress tests.

4. Results

4.1. Subject-Independent RAVDESS Validation

The revised evaluation was designed to distinguish exact pair matching from probabilistic multimodal consistency. Results are reported in two complementary ways: whether a model can discriminate genuine face–body pairs from mismatched pairs, and whether the learned consistency score remains meaningful under subject-independent splits, modality-disruption controls, identity-leakage analysis, cross-dataset transfer, and counterfactual conditions. Unless otherwise stated, reported values are ROC–AUC scores averaged over three random seeds.

Subject-independent validation on the expanded RAVDESS cache showed a clear difference between direct pair matching and prototype-based probabilistic agreement estimation. The no-prototype DirectCosine baseline ($K = 0$) achieved the strongest ordinary exact pair-matching performance, with ROC–AUC = 0.943 and PR–AUC = 0.917. Among PMCI variants, PMCI_K32 achieved the strongest standard pair-matching separation, with ROC–AUC = 0.887, followed by PMCI_K8 and PMCI_K16. These results support the intended interpretation of PMCI as an auxiliary diagnostic agreement index rather than as a replacement for direct embedding similarity. Table 2 summarizes the expanded RAVDESS standard pair-matching results.

Because the expanded RAVDESS cache contains four temporal crops per source video, we performed an additional control to test whether the standard pair-matching pattern was driven only by repeated windows from the same videos. In this control, one temporal window was randomly selected from each source video, and the standard pair-matching evaluation was repeated 100 times. The main pattern remained stable. DirectCosine_K0 achieved ROC–AUC = 0.942 plus-minus 0.009. PMCI_K16 achieved ROC–AUC = 0.807 plus-minus 0.016, and PMCI_K32 achieved ROC–AUC = 0.893 plus-minus 0.012. The standard pair-matching performance of DirectCosine_K0 and the PMCI variants is visualized in Figure 2. Thus, although the expanded cache increases the number of evaluated windows, the standard pair-matching pattern is not explained only by repeated temporal crops from the same source videos. The one-window-per-source-video control was used as a

source-video-aware robustness check rather than as a separate model comparison. Its purpose was to test whether the expanded RAVDESS cache artificially inflated the standard pair-matching results through repeated temporal crops from the same videos. The results are summarized in Table 3.

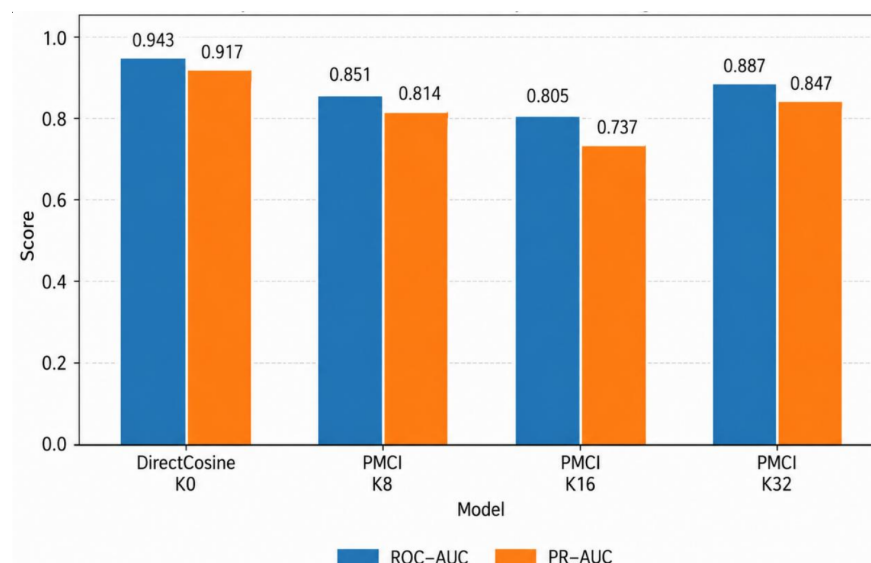


Figure 2. Subject-independent RAVDESS validation: ROC-AUC scores for the direct cosine baseline and PMCI variants across mismatch conditions.

Table 2. Expanded RAVDESS standard pair-matching metrics.

Model	ROC-AUC	PR-AUC	Positive Mean	Negative Mean
DirectCosine_K0	0.943	0.917	0.823	0.496
PMCI_K8	0.851	0.814	0.715	0.365
PMCI_K16	0.805	0.737	0.743	0.409
PMCI_K32	0.887	0.847	0.823	0.397

Table 3. One-window-per-source-video control across 100 repeats.

Model	ROC-AUC Mean	ROC-AUC Std
DirectCosine_K0	0.942	0.009
PMCI_K16	0.807	0.016
PMCI_K32	0.893	0.012

4.2. Modality-Disruption Controls

To verify that the evaluated scores respond to artificial disruption of the facial-gestural relationship, we performed additional modality-disruption controls on the expanded RAVDESS test split. These controls compare genuine face-body pairs against deliberately disrupted counterparts, including random body replacement, temporal reversal, and modality-zeroing conditions. Therefore, the values in this section should be interpreted as disruption-detection controls rather than as the main expanded RAVDESS pair-matching ranking reported in Table 2. The results show that random-body replacement produced strong separation for all evaluated scores. DirectCosine_K0 achieved ROC-AUC = 0.942, while PMCI_K8, PMCI_K16, and PMCI_K32 achieved ROC-AUC values of 0.838, 0.800, and 0.893, respectively. Modality-zeroing controls were also separable from genuine pairs, especially for DirectCosine_K0 and PMCI_K32. Temporal reversal was more difficult:

face-only and body-only reversal produced weaker separation, and reversing both streams was close to chance for DirectCosine_K0, PMCI_K16, and PMCI_K32. This indicates that the disruption controls do not simply measure model capacity, but capture different degrees of artificial mismatch severity. These results support the interpretation that PMCI is sensitive to relational changes between facial and gestural streams, while also confirming that not all perturbations are equally informative. In particular, random body replacement and modality-zeroing create strong artificial disruptions, whereas temporal reversal can preserve part of the short-window semantic structure. The results are summarized in Table 4.

Table 4. Expanded RAVDESS disruption-control evaluation (ROC–AUC; true pairs versus disrupted counterparts).

Model	Normal Pair Matching	Random Body	Face Zero	Body Zero	Both Time-Reversed
DirectCosine_K0	0.943	0.942	0.980	0.985	0.480
PMCI_K8	0.851	0.838	0.814	0.794	0.539
PMCI_K16	0.805	0.800	0.819	0.633	0.501
PMCI_K32	0.887	0.893	0.879	0.840	0.484

These disruption-control results should not be interpreted as showing that the model performs better when a modality is removed. Rather, they show that genuine face–body pairs are separable from artificially disrupted counterparts. The strongest separation was observed for random-body replacement and modality-zeroing conditions, whereas both-stream temporal reversal was close to chance for several models. This supports the interpretation that PMCI is sensitive to relational disruption, while also showing that short-window temporal reversal is a weaker perturbation than random replacement or complete modality removal. The same disruption-control pattern is visualized in Figure 3.

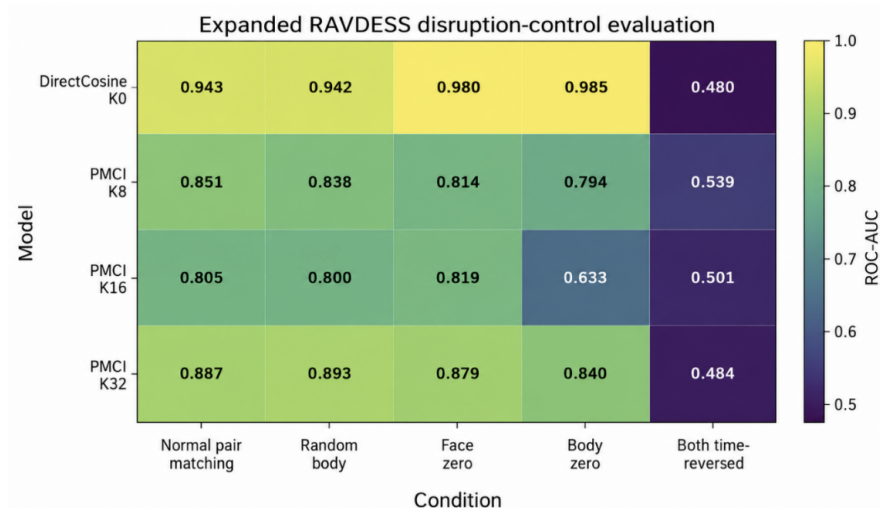


Figure 3. Expanded RAVDESS disruption-control evaluation.

4.3. MELD In-the-Wild Validation

MELD was used as a noisier in-the-wild conversational validation setting [30]. Unlike RAVDESS, which contains controlled acted emotional expressions, MELD includes shorter conversational clips, more variable scene composition, less consistent body visibility, and less controlled affective expression. Under within-dataset training and testing on MELD, PMCI $K = 8$ achieved 0.755 ± 0.007 AUC for cross-video mismatches and 0.750 ± 0.012 for shuffled mismatches, indicating that the proposed consistency score can still separate

genuine and mismatched face–body pairs in a noisy conversational setting. However, same-video and temporal-shift conditions were more difficult, with AUC values of 0.641 ± 0.086 and 0.627 ± 0.073 , respectively. Results are reported in Table 5.

Table 5. MELD within-dataset validation using PMCI $K = 8$ (ROC–AUC, mean $\pm$ std).

Mismatch Condition	AUC
Cross-video	0.755 ± 0.007
Same-video	0.641 ± 0.086
Temporal-shift	0.627 ± 0.073
Shuffle	0.750 ± 0.012

The performance of PMCI on MELD should not be regarded as conclusive cross-domain generalization, but as preliminary filtered in-the-wild validation. These results suggest that PMCI can remain partially informative in selected MELD conditions, especially for cross-video and shuffled mismatches. However, the weaker same-video and temporal-shift results indicate that its reliability decreases under noisier naturalistic visual conditions. This observation further reinforces the reading of PMCI as an indicator of brief-window, distribution-level consensus, rather than as a robust temporal-alignment model or a general-purpose emotion classifier. A better representation of the amount of temporal slippage would require additional modalities such as speech and text, more context, or explicit sequence-alignment modules.

4.4. Cross-Dataset Transfer and Adaptation

Cross-dataset transfer was evaluated by training on RAVDESS [29] and testing on MELD [30]. The zero-shot transfer result was close to chance: with 0% MELD target-domain adaptation, PMCI $K = 8$ achieved only 0.501 AUC for cross-video mismatch detection and 0.503 AUC for shuffled mismatches. This result indicates a substantial domain gap between controlled acted-emotion videos and noisy in-the-wild conversational clips. It reinforces that PMCI should not be interpreted as a universally domain-invariant consistency metric; domain-aware adaptation remains necessary.

Target-domain adaptation improved performance modestly. With 10%, 20%, 50%, and 100% MELD adaptation, cross-video AUC increased to 0.578, 0.573, 0.582, and 0.596, respectively. Shuffle AUC similarly increased from 0.503 in the zero-shot setting to 0.593 with full MELD adaptation. However, the improvement saturated early and did not approach within-domain MELD performance. The adaptation curve is demonstrated in Figure 4 and Table 6.

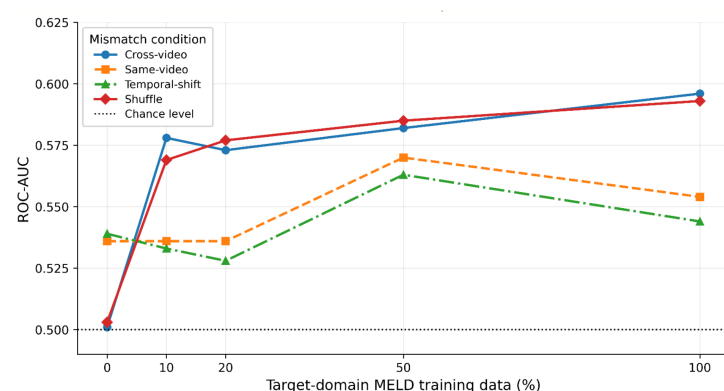


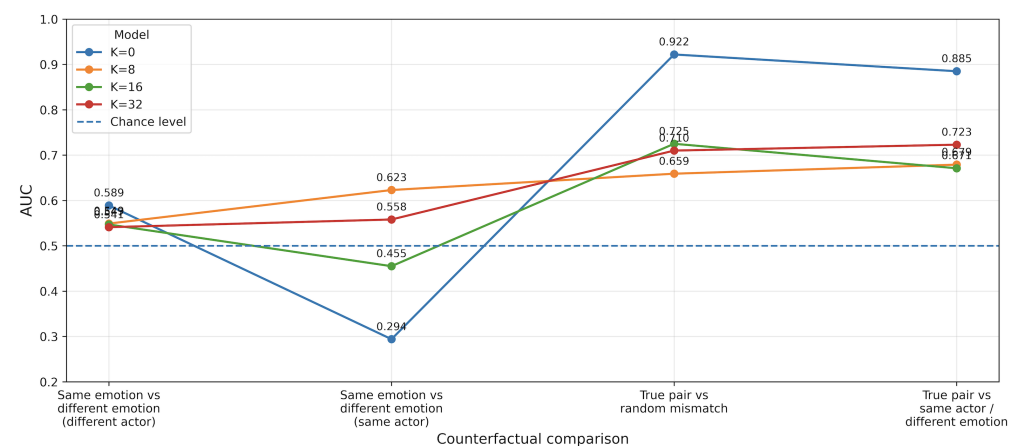
Figure 4. RAVDESS-to-MELD adaptation curve: cross-video and shuffled AUC as a function of the fraction of MELD target-domain training data used.

Table 6. RAVDESS-to-MELD adaptation curve using PMCI $K = 8$ (ROC–AUC).

Target-Domain MELD Data	Cross-Video	Same-Video	Temporal-Shift	Shuffle
0% MELD	0.501	0.536	0.539	0.503
10% MELD	0.578	0.536	0.533	0.569
20% MELD	0.573	0.536	0.528	0.577
50% MELD	0.582	0.570	0.563	0.585
100% MELD	0.596	0.554	0.544	0.593

4.5. Balanced Counterfactual Evaluation

To ensure that PMCI functions as a semantic consistency measure rather than merely as a video-pair detector, a balanced counterfactual test was conducted. In exact true-pair discrimination, direct cosine was the strongest baseline, achieving 0.922 ± 0.019 AUC for true pairs versus random mismatches and 0.885 ± 0.025 for true pairs versus same-actor/different-emotion pairs. In contrast, for the semantic identity-conflict condition (same emotion, different actor versus same actor, different emotion), direct cosine demonstrated an AUC of 0.294 ± 0.067 , significantly below chance. PMCI with $K = 8$ achieved 0.623 ± 0.036 AUC on the same task, an improvement of 0.329. Thus, while direct cosine is highly effective for exact pairwise discrimination, its efficacy diminishes when actor identity and affective compatibility are juxtaposed; PMCI retains the capacity to assess affective compatibility in these counterfactual scenarios. Because this test depends on actor/emotion metadata, it should be regarded as an additional line of diagnostic evidence rather than the primary validation of the proposed method's superiority. Results are shown in Figure 5 and Table 7.

**Figure 5.** Balanced counterfactual evaluation: ROC–AUC for direct cosine and PMCI variants across counterfactual pair types.**Table 7.** Balanced counterfactual evaluation (ROC–AUC, mean $\pm$ std).

Counterfactual Comparison	Direct Cosine $K = 0$	PMCI $K = 8$	PMCI $K = 16$	PMCI $K = 32$
Same emotion/different actor vs. different emotion/different actor	0.589 ± 0.061	0.549 ± 0.013	0.547 ± 0.033	0.541 ± 0.055
Same emotion/different actor vs. same actor/different emotion	0.294 ± 0.067	0.623 ± 0.036	0.455 ± 0.111	0.558 ± 0.101
True pair vs. random mismatch	0.922 ± 0.019	0.659 ± 0.075	0.725 ± 0.034	0.710 ± 0.007
True pair vs. same actor/different emotion	0.885 ± 0.025	0.679 ± 0.074	0.671 ± 0.015	0.723 ± 0.023

4.6. Identity-Leakage Analysis

The actor-level leakage analysis shows that although prototype-based PMCI removes some identity information, traces remain detectable. Leakage was highest with direct cosine: a linear identity classifier achieved 0.891 ± 0.002 balanced accuracy on concatenated face–body embeddings. With PMCI, identity-prediction accuracy fell to 0.713 ± 0.020 , 0.691 ± 0.074 , and 0.618 ± 0.067 for $K = 8, 16$, and 32 , respectively. Prototype-based consistency modelling can therefore reduce reliance on actor information without eliminating it. PMCI is consequently treated as a less identity-dominated diagnostic consistency measure rather than an identity-invariant representation. Results are shown in Figure 6 and Table 8.

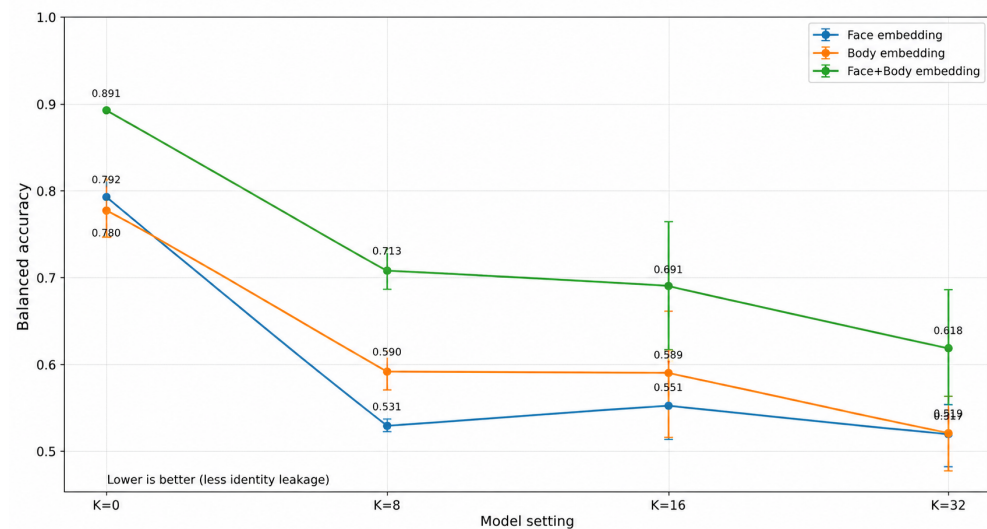


Figure 6. Actor-identity leakage analysis: balanced accuracy of a linear identity classifier on face, body, and face+body embeddings for each model variant.

Table 8. Actor-identity leakage analysis (balanced accuracy, mean $\pm$ std).

Model	Face Embedding	Body Embedding	Face + Body Embedding
Direct cosine, $K = 0$	0.792 ± 0.008	0.780 ± 0.033	0.891 ± 0.002
PMCI, $K = 8$	0.531 ± 0.006	0.590 ± 0.020	0.713 ± 0.020
PMCI, $K = 16$	0.551 ± 0.037	0.589 ± 0.073	0.691 ± 0.074
PMCI, $K = 32$	0.517 ± 0.036	0.519 ± 0.043	0.618 ± 0.067

Overall, the results show that PMCI should not be interpreted as a universal replacement for direct embedding similarity. Direct cosine remains stronger for exact pair discrimination. The contribution of PMCI is different: it provides a probabilistic, interpretable, and less identity-dominated consistency measure whose behavior remains meaningful under modality dropout, domain shift, and counterfactual controls. These findings support the use of PMCI as a diagnostic layer for multimodal affective consistency analysis rather than as a purely discriminative pair-matching model.

4.7. Additional Non-Redundancy and Reviewer-Oriented Stress Tests

To assess the non-redundancy of PMCI vis-a-vis embedding similarity, we conducted reviewer-oriented stress tests: hard-negative evaluation, shuffled controls, decile analysis, and grouped bootstrap. These tests were designed to probe PMCI in cases where basic embedding similarity has little discriminatory power. Across DirectCosine-mined hard negatives, DirectCosine was at or marginally below chance (ROC–AUC = 0.484), whereas

both PMCI_K16 and PMCI_K32 showed some degree of separation above chance (ROC-AUC = 0.572 and 0.569, respectively), and DirectCosine + PMCI_K32 had a score of ROC-AUC = 0.582. The magnitude of this increase is relatively small, although shuffled-control and permutation tests suggest it was a genuine, statistically stable increase and does not reflect the random contribution of another dimension. The grouped bootstrap across the source video for the Direct + PMCI_K32 model was in the range of [0.569, 0.596]. Decile analysis provides a more diagnostic look at this same increase. The two lowest PMCI_K32-score deciles were made up of no matched pairs, while the third decile had 94.4% mismatched. This suggests a possible explanation of low PMCI as a low-agreement signal rather than a generic emotion classifier. For completeness, we performed identity-shortcut controls by constructing hard negatives without the same source video as well as same actor. In this case, DirectCosine showed strong discrimination, suggesting that these hard cross-actor negatives are a different and unambiguous regime, not the same as DirectCosine-mined hard negatives. PMCI_K32 was still informative (ROC-AUC = 0.800), but this is an identity-shortcut check rather than primary evidence for non-redundancy.

4.8. Technical Checks

We ran additional tests to see if PMCI results might be due to arbitrary or unstable parameter selection. These were not intended as an alternative model or new architecture, but as additional analysis to determine sensitivity to initialization parameters, temperature parameters, anchor usage, and pose estimation. Our first test was on initialization of the latent anchors in the shared space. We used normal, uniform, Xavier, and orthogonal initialization; all produced very similar PMCI patterns and scores, suggesting that PMCI performance is not sensitive to a specific initialization. We also probed the temperature parameter in the joint-softmax over the shared anchors. Moderate values for the temperature parameter were sufficient to generate sensible, unimodal distributions, but larger temperature values led to broader distributions with less informative JS scores. In other words, the temperature parameter changes the sharpness rather than the identity of the modalities. Computational performance was also measured, and the total pipeline required 10.46 ms/frame for YOLOv8 pose extraction, and 0.021 ms/window for PMCI_K32. This suggests that PMCI calculation can be considered a low-cost operation once keypoints are already extracted. We also inspected cases with low facial/gestural overlap. These cases show a decreased PMCI, exactly as expected when pF and pB distribute the same anchor probability mass to different regions of anchor space and have low JS agreement. That is, PMCI is an agreement measure, and PMCI should not increase if the distributions do not overlap. The number of prototype anchors was also verified. In neither of the baseline K16 or K32 models did we encounter severe underutilization of the anchors. The anchor model was sufficient to support the interpretation of PMCI as an indicator of distributional compatibility rather than simply the probability of match to a prototype anchor. Lastly, we tested PMCI robustness to noise in pose estimation. We found that mild keypoint jitter had little impact on the score, whereas keypoint omission did impact performance, as expected for keypoint-based analysis only. This confirms an explicit limitation in PMCI, namely that PMCI requires high-quality keypoint annotation, and cannot make up for missing modalities such as speech, vocal prosody, facial texture, text, or appearance.

5. Discussion

The results of this study show that probabilistic multimodal consistency should be interpreted as a diagnostic measure of cross-modal affective agreement rather than as a universal replacement for direct embedding similarity. The strongest exact pair-matching performance was obtained by the no-prototype direct cosine baseline. However, this high

discriminative performance comes with a cost: the learned representations retain stronger actor-identity information.

The most significant additional finding is that when embedding-based similarity is ambiguous, PMCI becomes a useful discriminator. For DirectCosine-mined hard negatives, direct similarity was no longer informative, whereas PMCI-based gating recovered a discriminative signal, suggesting that PMCI contributes non-redundant prototype-distribution consistency information. The shuffled-PMCI control confirmed that this improvement was not explained by additional features or classifier capacity: it disappeared when PMCI was randomly shuffled and remained when real PMCI was used.

The main contribution of PMCI is a more controlled probabilistic formulation of multimodal consistency. By representing facial and gestural streams as modality-specific probability distributions over shared affective prototypes, PMCI provides an interpretable score of whether the two modalities support compatible affective interpretations. The identity-leakage analysis showed that prototype-based PMCI reduces actor-identity information compared with direct cosine, although it does not eliminate it completely. This is important for interpreting the results because a high pair-matching AUC alone does not demonstrate that the model measures affective consistency: such a score may partly arise from actor identity, individual motion style, recording conditions, or stable video-level shortcuts. In exact pair discrimination, a model may successfully determine that facial and bodily streams belong to the same video segment, but this success does not necessarily mean that it has captured the compatibility of the expressed affective state.

The present study therefore considers exact matching of face and gesture pairs and semantic affective consistency as related but distinct tasks. Direct similarity can be highly effective for determining whether facial and bodily streams originate from the same segment; however, such effectiveness may partially rely on identity-sensitive information. PMCI has a different goal: it evaluates whether the two modalities support a compatible affective interpretation at the level of probability distributions. Such framing is consistent with perceptual studies showing that facial and bodily affective signals are rapidly and contextually integrated in human emotion perception [6–8]. Meeren et al. reported early neural integration of facial and bodily emotional information, while Aviezer et al. showed that body cues can dominate facial cues under intense affective conditions [6,7]. Lecker et al. further demonstrated bidirectional contextual influence between facial and bodily expressions [8]. In this context, the present computational results support the view that face–body consistency is not merely a fusion problem, but a measurable cross-modal relation.

The findings indicate that PMCI tackles a distinct problem from established methodologies in multimodal affective computing. Fusion, attention-based alignment, and latent-representation matching improve emotion classification or downstream-task performance [14–16]. In contrast, PMCI asks whether facial and gestural streams have affectively compatible interpretations. This differs from asking whether a fused model improves recognition, because such a model may be dominated by one modality. PMCI directly addresses consistency between facial and gestural modalities.

The comparison with the direct cosine baseline demonstrates a clear trade-off. Direct cosine remains a stronger method for the exact determination of matching face–body pairs, especially under subject-independent validation on RAVDESS [29]. However, the identity leakage analysis shows that such efficiency is partly connected with actor-specific information. PMCI does not replace direct similarity; rather, it provides a more controlled probabilistic signal that is less dominated by identity features. Such behavior is especially noticeable in the balanced counterfactual setting, where PMCI with $K = 8$ showed better results in the identity–emotion conflict condition when comparing pairs with the same emotion but different actors against pairs with the same actor but different emotions.

The control experiments with modality disruption additionally confirm this interpretation. When facial information, body information, or the correct matching of the body stream was removed, the effectiveness dropped. This shows that the indicator depends specifically on the relationship between facial and gestural streams, and not on one separate modality alone. For a measure of multimodal consistency, such results are important.

The MELD results show that PMCI can remain informative in noisy in-the-wild conversational clips, but that the consistency signal is domain-sensitive [30]. Zero-shot transfer from RAVDESS to MELD was close to chance, and target-domain adaptation improved performance only modestly, indicating a substantial domain gap between controlled acted-emotion videos and unconstrained conversational scenes. Therefore, PMCI should not be claimed to generalize universally across datasets without domain-specific validation.

The use of Jensen–Shannon divergence remains appropriate for the proposed formulation because it compares full probability distributions rather than unconstrained latent vectors, and is symmetric and bounded [33]. However, the present results also show that this probabilistic interpretation should be evaluated together with shortcut controls.

Despite the limited number of independent source windows, the additional analyses demonstrate the stability of the PMCI effect under reviewer-oriented stress conditions. Most notably, PMCI-based gating recovers a discriminative signal in DirectCosine-mined hard negatives where direct embedding similarity is uninformative. Real PMCI also consistently outperformed shuffled PMCI, so the gain cannot be attributed solely to additional feature dimensions. The revised analysis therefore supports PMCI as a reliable, non-redundant diagnostic consistency indicator while acknowledging that it does not surpass direct similarity in routine exact pair matching.

A straightforward numerical comparison with existing multimodal emotion-recognition findings is not possible here, since most comparable studies report emotion-classification accuracy or F1-score on audio-visual and audio-textual test sets, while in this work we report the pose-based cross-modal agreement and ability to discriminate exact versus shuffled pairs. Therefore, the most comparable comparisons in this study are internal: DirectCosine and LateFusionMLP constitute traditional exact pair matching methods, and the shuffled, random, untrained, and disruption conditions are used to determine if PMCI provides additional diagnostic utility. The results should hence be interpreted as demonstrating an additional agreement index for multimodal emotion recognition rather than as a demonstration of best-in-class performance in the recognition of emotion from audiovisual or audio-only inputs.

To contextualize the magnitude of the reported values, it is important to compare them with prior multimodal emotion-recognition results only cautiously. Most related studies report emotion-classification accuracy or F1-score, whereas the present work reports ROC–AUC for face–body pair-consistency and cross-modal agreement discrimination. For example, a multimodal RAVDESS emotion-recognition study based on speech and facial information reported 80.08% accuracy under subject-wise cross-validation [36], while multimodal deep-learning work on affective physiological data reported 83.25% mean accuracy on DEAP [37]. These values are not directly interchangeable with the ROC–AUC values reported here, because the underlying tasks, modalities, and evaluation protocols differ. Nevertheless, they provide a useful scale: the DirectCosine_K0 result on expanded RAVDESS (ROC–AUC = 0.943) is consistent with the expectation that controlled RAVDESS-style settings allow strong discrimination, whereas PMCI variants (ROC–AUC = 0.805–0.887) remain below the strongest exact-matching baseline but within a meaningful diagnostic range. By contrast, the hard-negative Direct+PMCI_K32 result (ROC–AUC = 0.582) should be interpreted separately, because it reflects a deliberately difficult DirectCosine-mined diagnostic setting rather than ordinary emotion classifica-

tion or standard pair matching. Therefore, the comparison with the literature supports the intended interpretation of PMCI as an auxiliary agreement index rather than as a general-purpose emotion-recognition model.

The present study also had other limitations. First, although the updated experiments address, they do not eliminate the limitation of sample non-independence, as 11,520 pose-derived 12-frame windows were created from all 24 actors in the RAVDESS database by extracting four temporal crops per video from each actor, and thus multiple windows from a video cannot be considered independent behavioral samples. As such, we used a one-window-per-video control and a source video/actor grouped bootstrap for validation. These analyses provide some additional support for the stability of the main pattern, but additional studies, with more independent source videos, more variation in behavior, additional modalities, and naturalistic datasets, are needed for further validation.

Second, direct embedding similarity remains stronger for ordinary exact pair matching, so PMCI should not be presented as a universally superior classifier. Third, actor identity information is reduced but not eliminated. Fourth, zero-shot cross-dataset transfer remains weak. Fifth, the temporal-shift protocol should be interpreted as a short-window semantic perturbation rather than as a test of fine-grained synchrony. Finally, the number of prototype anchors is selected empirically, and future work should evaluate richer modality sets, stronger cross-domain adaptation methods, and more explicit semantic supervision.

Our final limitation concerns modality scope. This version of PMCI only considers pose-based facial and body keypoint data and does not take into account modality information including speech, text, vocal prosody, facial details, or full RGB video appearance. This choice limits the extent of identity and background shortcuts that can develop, but it also constrains the types of conclusions that are possible. In that sense, the PMCI that is evaluated in this paper can most accurately be described as a facial-gestural agreement measure based on pose, rather than a full multimodal emotion understanding system. Finally, another limitation concerns temporal modeling. Our temporal modeling is restricted to local differences in coordinates across short time windows in the present study, rather than explicit modeling of affective dynamics, delayed cross-modal response, or complex temporal misalignment over longer time intervals. Future studies can build on our PMCI approach with more explicit temporal alignment methods, longer sequence modeling, and more diverse multimodal channels.

Overall, the results suggest that PMCI is most useful as a diagnostic layer for multimodal affective consistency analysis. It should complement, rather than replace, direct embedding similarity. Direct cosine is better suited for exact pair matching, whereas PMCI provides a probabilistic, interpretable, and less identity-dominated measure of whether facial and gestural streams support compatible affective interpretations. Recent in-the-wild studies of co-speech gestures further support the need for realistic body-dynamics modeling in natural interaction settings [38].

From a theoretical standpoint, PMCI operates distinctly from fusion models, attention-based alignment architectures, straightforward cosine similarity, and LateFusionMLP. Fusion and alignment techniques primarily optimise downstream recognition or representation alignment, while DirectCosine and LateFusionMLP serve as strong pair-matching baselines. PMCI uniquely assesses whether facial and gestural modalities yield mutually consistent affective interpretations through prototype probability distributions. The additional reviewer-oriented evaluations are summarised in Table 9.

Although our revised experiments do not wholly resolve the sample independence issue, they do lessen its impact. For instance, we augmented the RAVDESS cache to incorporate 11,520 pose-derived 12-frame windows from each of 24 RAVDESS actors, using four temporal crops of each source RAVDESS video. Still, some windows in this cache could

originate from the same source video and, as such, are not fully independent behavioral samples. For this reason, we included one-window-per-source-video control analyses and analyses involving grouped bootstrap resampling by source video/actor. While our analyses suggest a stable main pattern, broader validation will still be necessary via larger naturalistic datasets with more source videos, richer gesture variability, and additional modalities. PMCI's present implementation also remains pose-based. While we leverage facial and body keypoint geometry along with coordinate differences over short temporal windows, our implementation does not incorporate speech, text, vocal prosody, facial texture, or full-color RGB appearance. Consequently, the current findings support PMCI's validity as a pose-derived facial–gestural agreement index rather than as a multimodal emotion-understanding system as a whole. Cross-domain transfer is also limited, particularly for MELD. In zero-shot mode, RAVDESS-to-MELD transfer nearly performs at chance, and even the best target domain adaptation results only yield modest gains in performance for PMCI. Thus, we should not conceive of the current PMCI approach as a domain-invariant universal consistency measure. Another limitation concerns the semantic interpretability of the learned prototype anchors. While the prototype layer enables the definition of a structured probability space for comparing facial and gestural evidence, our learned prototype anchors should not be considered verified psychological emotion categories. As we mentioned previously, the hyperparameter K defines the granularity of the latent agreement space rather than the number of distinct emotions in our model. Additional work should assess prototype interpretability through external annotations, expert ratings, cluster visualizations, and cross-dataset semantic validation. The MELD results should be viewed with more skepticism. Given the pose-only nature of the current PMCI implementation's performance against naturalistic image noise, limited body visibility, and quality of the pose-derived sample windows, we should only consider MELD a prefiltered in-the-wild test rather than proof of large-scale cross-domain performance robustness.

Table 9. Summary of additional reviewer-oriented PMCI evaluations.

Evaluation	Results	Interpretation
Expanded RAVDESS standard pair matching	DirectCosine_K0 = 0.943; PMCI_K32 = 0.887	Direct similarity remains stronger for ordinary pair matching
One-window-per-source-video control	PMCI_K32 = 0.893 $\pm$ 0.012	Pattern is not explained only by repeated crops
DirectCosine-mined hard negatives	Direct = 0.484; Direct+PMCI_K32 = 0.582	PMCI provides modest but stable diagnostic gain
Decile analysis	Two lowest K32 deciles contain 100% mismatches	Low PMCI works as a low-agreement warning signal
Strict identity-excluded hard negatives	PMCI_K32 = 0.800	PMCI remains informative under identity- shortcut control

6. Conclusions

This study introduces a probabilistic approach for assessing affective consistency between facial and gesture-based cues in interactive environments. Rather than framing multimodal analysis strictly as recognition or feature fusion, PMCI evaluates the extent to which affective interpretations derived from each modality align to establish a coherent cross-modal affective condition. Facial and bodily streams are processed separately and mapped to probability distributions over a shared set of latent affective-agreement anchors, and their coherence is quantified using Jensen–Shannon divergence. The experimental analysis indicates that PMCI should be regarded as a diagnostic consistency metric rather than as a complete substitute for direct embedding similarity.

Although DirectCosine and LateFusionMLP remain stronger baselines for standard exact face–body correspondence, the reviewer-oriented evaluation shows that PMCI provides a provided complementary diagnostic information under more rigorous test scenarios. PMCI-based gates improved discrimination in a comprehensive balanced pair bank and recovered signal from DirectCosine-mined hard negatives after conventional embedding similarity had become unreliable. The shuffled-PMCI validation further supports the methodology: performed above shuffled-control variants, confirming that the improvement cannot be attributed simply to increased feature dimensionality or classifier capacity. Moreover, $1 - \text{PMCI}$ serves as an information-geometric indicator of inconsistency. Taken together, these findings support PMCI as a bounded, distribution-level assessment of cross-modal disagreement.

PMCI nevertheless remains domain-sensitive. Zero-shot transfer from RAVDESS to MELD was negligible, and target-domain adaptation yielded only modest gains. PMCI therefore should not be claimed to apply universally across datasets without appropriate domain-specific validation. Its merit lies not in universality or superiority over embedding similarity, but in providing an additional diagnostic instrument for determining whether distinct modes of expression yield mutually congruent affective cues. Such a tool is particularly important in interactive contexts, where confusion, emotional suppression, hesitation, or transitional affective states may convey more salient information than a single emotion label.

Future research should examine PMCI on larger and more heterogeneous face–body video collections, retain source-video and emotion-label information throughout preprocessing, and incorporate a wider range of modalities. Further work should develop stronger domain-adaptation methods, PMCI variants that capture fine-grained temporal synchrony (including lag-sensitive or trajectory-based formulations), and real-time implementations for uncertainty-aware affective computing. The present conclusions are limited to facial and body-pose features and short-term motion patterns; they do not extend to comprehensive affective behaviour inferred from raw RGB video.

Overall, we characterize PMCI as a prototype-based probabilistic diagnostic for assessing cross-modal affective consistency. The findings from the augmented RAVDESS experiments and the additional controls that account for the source video suggest that PMCI does not serve as a proxy for embedding similarity in the usual setting of matching pairs; however, it does offer a small yet statistically reliable agreement signal in the setting of hard negatives and low agreement. Going forward, it would be useful to examine whether PMCI generalizes to bigger, naturalistic datasets, to settings with more diverse modalities, and to longer sequences.

Author Contributions: Conceptualization, Y.S. and S.K.; methodology, Y.S.; software, B.S.; validation, U.B., S.K. and A.B.; formal analysis, E.D. and U.B.; investigation, Y.S. and A.Y.; resources, S.K. and E.D.; data curation, B.S.; writing—original draft preparation, Y.S. and A.Y.; writing—review and editing, S.K. and A.B.; supervision, S.K. and A.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research has been funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. BR24992875).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The source code and experimental materials used in this study, including preprocessing, training, evaluation, result files, and figure-generation scripts, are available at: <https://github.com/ernar65019920/PMC-of-Facial-and-Gestural-Affective-Cues> (accessed on 16 July 2026).

Acknowledgments: The authors thank the creators of the RAVDESS and MELD datasets for making their data publicly available.

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

Abbreviations

The following abbreviations are used in this manuscript:

PMCI	Probabilistic Multimodal Consistency Index
EC	Emotion Congruence (operational consistency score)
GRU	Gated Recurrent Unit
JSD	Jensen–Shannon Divergence
AUC	Area Under the ROC Curve
RAVDESS	Ryerson Audio-Visual Database of Emotional Speech and Song
MELD	Multimodal EmotionLines Dataset
RGB	Red–Green–Blue

References

1. Darwin, C. *The Expression of the Emotions in Man and Animals*; John Murray: London, UK, 1872.
2. Ekman, P.; Friesen, W.V. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica* **1969**, *1*, 49–98. [\[CrossRef\]](#)
3. Plutchik, R. *Emotion: A Psychoevolutionary Synthesis*; Harper & Row: New York, NY, USA, 1980.
4. Russell, J.A. A circumplex model of affect. *J. Personal. Soc. Psychol.* **1980**, *39*, 1161–1178. [\[CrossRef\]](#)
5. Scherer, K.R. What are emotions? And how can they be measured? *Soc. Sci. Inf.* **2005**, *44*, 695–729. [\[CrossRef\]](#)
6. Meeren, H.K.M.; van Heijnsbergen, C.C.R.J.; de Gelder, B. Rapid perceptual integration of facial expression and emotional body language. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 16518–16523. [\[CrossRef\]](#) [\[PubMed\]](#)
7. Aviezer, H.; Trope, Y.; Todorov, A. Body cues, not facial expressions, discriminate between intense positive and negative emotions. *Science* **2012**, *338*, 1225–1229. [\[CrossRef\]](#) [\[PubMed\]](#)
8. Lecker, M.; Dotsch, R.; Bijlstra, G.; Aviezer, H. Bidirectional contextual influence between faces and bodies in emotion perception. *Emotion* **2020**, *20*, 1154–1164. [\[CrossRef\]](#) [\[PubMed\]](#)
9. Barros, P.; Wermter, S. Developing crossmodal expression recognition based on a deep neural model. *Adapt. Behav.* **2016**, *24*, 373–396. [\[CrossRef\]](#) [\[PubMed\]](#)
10. Hu, G.; Kollias, D.; Yang, X. Grounding emotion recognition with visual prototypes: VEGA. In *Proceedings of ACM Multimedia (ACM MM)*; ACM: New York, NY, USA, 2025.
11. Chhatre, K.; Guarese, R.; Matvienko, A.; Peters, C. Evaluating speech and video models for face-body congruence. In *Proceedings of the ACM International Conference on 3D Web Technology (I3D) Companion*; ACM: New York, NY, USA, 2025.
12. Puffet, A.-S.; Rigoulot, S. The role of cognitive load in automatic integration of emotional information from face and body. *Sci. Rep.* **2025**, *15*, 28184. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Reisenzein, R.; Studtmann, M.; Horstmann, G. Coherence between emotion and facial expression: Evidence from laboratory experiments. *Emotion* **2013**, *13*, 737–753.
14. Poria, S.; Cambria, E.; Bajpai, R.; Hussain, A. A review of affective computing: From unimodal analysis to multimodal fusion. *Inf. Fusion* **2017**, *37*, 98–125. [\[CrossRef\]](#)
15. Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; Morency, L.-P. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Copenhagen, Denmark, 2017; pp. 1103–1114. [\[CrossRef\]](#)
16. Tsai, Y.-H.H.; Bai, S.; Liang, P.P.; Kolter, J.Z.; Morency, L.-P.; Salakhutdinov, R. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Florence, Italy, 2019; pp. 6558–6569. [\[CrossRef\]](#) [\[PubMed\]](#)
17. Ma, Z.M.; Zhang, W.J.; Ma, W.Y. Semantic measure of fuzzy data in extended possibility-based fuzzy relational databases. *Int. J. Intell. Syst.* **2000**, *15*, 705–716. [\[CrossRef\]](#)
18. Agarwala, S.; Anagawadi, A.; Guddeti, R.M.R. Detecting semantic similarity of documents using natural language processing. *Procedia Comput. Sci.* **2021**, *189*, 128–135. [\[CrossRef\]](#)
19. Akila, D.; Jayakumar, C. Semantic similarity—A review of approaches and metrics. *Int. J. Appl. Eng. Res.* **2014**, *9*, 27581–27600. [\[CrossRef\]](#)

20. Liu, X.; Ghorpade, A.; Tu, Y.L.; Zhang, W.J. A novel approach to probability distribution aggregation. *Inf. Sci.* **2012**, *188*, 269–275. [\[CrossRef\]](#)
21. Kleinsmith, A.; Bianchi-Berthouze, N. Affective body expression perception and recognition: A survey. *IEEE Trans. Affect. Comput.* **2013**, *4*, 15–33. [\[CrossRef\]](#)
22. Poyo Solanas, M.; Vaessen, M.; de Gelder, B. The role of kinematics and posture in body expression perception. *Soc. Cogn. Affect. Neurosci.* **2020**, *15*, 853–867.
23. Cao, Z.; Simon, T.; Wei, S.-E.; Sheikh, Y. Realtime multi-person 2D pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Honolulu, HI, USA, 2017.
24. de Gelder, B. Towards the neurobiology of emotional body language. *Nat. Rev. Neurosci.* **2006**, *7*, 242–249. [\[CrossRef\]](#) [\[PubMed\]](#)
25. Chung, J.S.; Zisserman, A. Out of time: Automated lip sync in the wild. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*; Springer: Cham, Switzerland, 2016.
26. Koelstra, A.; Muhl, C.; Soleymani, M.; Lee, J.-S.; Yazdani, A.; Ebrahimi, T.; Pun, T.; Nijholt, A.; Patras, I. DEAP: A database for emotion analysis using physiological signals. *IEEE Trans. Affect. Comput.* **2012**, *3*, 18–31. [\[CrossRef\]](#)
27. Zeng, H.; Wang, X.; Wu, A.; Wang, Y.; Li, Q.; Endert, A.; Qu, H. EmoCo: Visual analysis of emotion coherence in presentation videos. *IEEE Trans. Vis. Comput. Graph.* **2020**, *26*, 928–937.
28. Liu, H.; Zhu, Z.; Iwamoto, N.; Peng, Y.; Li, Z.; Zhou, Y.; Bozkurt, E.; Zheng, B. BEAT: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In *Computer Vision—ECCV 2022*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2022; Volume 13667, pp. 612–630. [\[CrossRef\]](#)
29. Livingstone, S.R.; Russo, F.A. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLoS ONE* **2018**, *13*, e0196391. [\[CrossRef\]](#) [\[PubMed\]](#)
30. Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; Mihalcea, R. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, 28 July–2 August 2019; pp. 527–536. [\[CrossRef\]](#)
31. Jocher, G.; Chaurasia, A.; Qiu, J. *Ultralytics YOLOv8*; Ultralytics: Los Angeles, CA, USA, 2023.
32. Cho, K.; van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*; Association for Computational Linguistics: Doha, Qatar, 2014; pp. 1724–1734. [\[CrossRef\]](#)
33. Lin, J. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory* **1991**, *37*, 145–151. [\[CrossRef\]](#)
34. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv* **2019**, arXiv:1711.05101. [\[CrossRef\]](#)
35. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874. [\[CrossRef\]](#)
36. Luna-Jiménez, C.; Kleinlein, R.; Griol, D.; Callejas, Z.; Montero, J.M.; Fernández-Martínez, F. Multimodal emotion recognition on RAVDESS dataset using transfer learning. *Sensors* **2021**, *21*, 7665. [\[CrossRef\]](#) [\[PubMed\]](#)
37. Liu, W.; Zheng, W.-L.; Lu, B.-L. Multimodal emotion recognition using multimodal deep learning. *arXiv* **2016**, arXiv:1602.08225. [\[CrossRef\]](#)
38. Hegde, S.B.; Prajwal, K.R.; Kwon, T.; Zisserman, A. Understanding co-speech gestures in-the-wild. *arXiv* **2025**, arXiv:2503.22668. [\[CrossRef\]](#)

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.