big data and cognitive computing

MDPI

Article

Unified Visual Synchrony: A Framework for Face–Gesture Coherence in Multimodal Human–AI Interaction

Saule Kudubayeva [1,2], Yernar Seksenbayev [2,*], Aigerim Yerimbetova [1,3,*], Elmira Daiyrbayeva [1,4], Bakzhan Sakenov [1], Duman Telman [1,5] and Mussa Turdalyuly [1,3]

1 Institute of Information and Computational Technologies CS MSHE RK, Almaty 050000, Kazakhstan; saule.kudubayeva@gmail.com (S.K.); e.daiyrbayeva@satbayev.university (E.D.); sbakzhan22@gmail.com (B.S.); telmand28@gmail.com (D.T.); m.turdalyuly@gmail.com (M.T.)
2 Faculty of Digital Sciences and Artificial Intelligence, L. N. Gumilyov Eurasian National University, Astana 010008, Kazakhstan
3 School of Engineering and Information Technology, META University, Almaty 050012, Kazakhstan
4 Department of Software Engineering, Satbayev University, Almaty 050010, Kazakhstan
5 School of Information Technologies and Applied Mathematics, SDU University, Kaskelen 040901, Kazakhstan
* Correspondence: simpleuser6501@gmail.com (Y.S.); aigerian8888@gmail.com (A.Y.)

Abstract

Multimodal human–AI systems generally consider facial expressions and body motions as separate input streams, leading to disjointed interpretations and diminished emotional coherence. To overcome this issue, we offer the Engagement-Safe Expressive Alignment (ESEA) paradigm and the Unified Visual Synchrony (UVS) framework as its computational implementation. UVS models the coherence between facial expressions and gestures, offering an interpretable visual synchrony signal that can function as adaptive feedback in human–AI interactions. The framework's key component is the Consistency Index for Affective Synchrony ($CIAS$), which correlates brief visual segments with scalar synchrony scores through a common latent representation. Facial and gestural signals are processed by modality-specific projection networks into a unified latent space, and $CIAS$ is derived from the similarity and short-term temporal consistency of these latent trajectories. The synchrony index is regarded as an estimation of affective visual coherence within the ESEA paradigm. We formalize the UVS/$CIAS$ framework and conduct a comparative experimental evaluation utilizing matched and mismatched face–gesture segments derived from rendered dialog footage. Utilizing ROC analysis, score distribution comparisons, temporal visualizations, and negative control tests, we illustrate that $CIAS$ effectively captures structured face–gesture alignment that surpasses similarity-based baselines, while also delivering a persistent, time-resolved synchronization signal. These findings establish $CIAS$ as a principled and interpretable feedback signal for future affect-aware, engagement-focused multimodal agents.

Keywords: multimodal fusion; synchrony; affective computing; gesture analysis; facial expression; human–AI interaction; adaptive systems

Check for updates

Academic Editor: Fabrizio Messina

Received: 15 January 2026
Revised: 28 February 2026
Accepted: 9 March 2026
Published: 12 March 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) license.

1. Introduction

Human communication is inherently multimodal and is realized through coordinated patterns of speech, hand gestures and facial expressions. Psychological and affective computing studies indicate that emotions are encoded in joint configurations of verbal and non-verbal behavior, so a unified analysis of facial expressions and body gestures

is required to model expressive communication in a reliable way [1]. Non-verbal cues such as the qualitative form of gestures, the synchrony between facial expressions and body posture, and eye gaze contribute both to affect expression and to the regulation of turn-taking and feedback in interaction [2].

Most current human–AI interaction (HAI) systems still remain limited to unimodal processing of visual behavior. Emotion-aware interfaces typically operate either on facial expressions or on body gestures in isolation, without explicitly modelling cross-modal synchrony, which is the temporal and semantic alignment between expressive channels [3]. As a result, the system cannot distinguish between configurations where a user's smile and an open hand gesture form a coherent affective act and configurations where facial and bodily cues conflict. This lack of joint modelling can lead to fragmented or inappropriate system responses when the user's signals are not interpreted as a single multimodal pattern.

In this research paper, we introduce a computational framework for unified visual synchrony that integrates gestures and facial expressions into a single interpretable representation. The framework is formulated within the Engagement-Safe Expressive Alignment (ESEA) paradigm, which treats timing and expressivity of multimodal behavior as a controllable resource constrained by perceived naturalness [4]. The motivation for this framework arises from three limitations of existing approaches. First, current emotion-recognition and gesture-recognition models are usually trained and deployed as separate pipelines, which prevents the system from modelling conditional dependencies between facial and gestural cues and from capturing whether two channels jointly realize one affective act or indicate a conflict [5]. Second, temporal coherence between face and body signals is rarely used as an explicit quantitative measure of expressiveness; while work in animation and virtual agents has highlighted the importance of perceived face–body congruence for realism, there is no widely adopted synchrony metric for live interaction that operates at the level of continuous behavior [6,7]. Third, adaptive HAI systems lack continuous and interpretable feedback signals that reflect the dynamics of the user's state: a single frame-level emotion label (for example, happy or angry) is insufficient for fine-grained adaptation, whereas a time-varying synchrony score that reflects how consistently the user's visual channels agree could provide a natural control signal for modulating the agent's behavior, such as its prosodic profile, pacing or degree of expressed empathy [8].

We brainstorm that synchrony between facial and gestural cues can be modelled as a latent signal that reflects both affective consistency and interactional engagement. Intuitively, when a person's face and body express similar affective content, for example both indicating enthusiasm, the overall pattern is coherent, whereas mismatched cues such as a smiling face combined with a closed, defensive posture may indicate ambivalence or discomfort [9]. This notion is compatible with dimensional and discrete views of emotion. For instance, Plutchik's wheel of emotions organizes primary emotions in opposing pairs (Joy–Sadness, Anger–Fear, Trust–Disgust, Surprise–Anticipation) (see Figure 1) and assigns intensities to each region [10]. If facial expression and body posture map to the same or neighboring regions on this wheel, we treat the visual cues as affectively synchronized; if they map to distant or opposing regions, affective synchrony is low. By mapping facial and gestural features into a common affective space, the proposed method aims to quantify how consistently both channels reflect a single emotional state over time.

This work makes three contributions within this context. To enhance the suggested paradigm with empirical evidence, we do a comparative experimental study incorporating similarity-based baselines, temporal visualizations, and negative control trials. This analysis aims to illustrate that the suggested Consistency Index for Affective Synchrony ($CIAS$) encapsulates organized face–gesture alignment beyond mere instantaneous feature similarity and offers a comprehensible, time-resolved synchrony signal. First, we articu-

late a multimodal architecture for joint encoding of facial and gestural dynamics, where modality-specific encoders project facial expression features and body gesture features into a shared latent space with temporal alignment, so that synchrony between modalities can be computed directly at the level of short temporal windows. Secondly, we formally define the $CIAS$, a continuous metric of face-body synchrony over time, which is based on the similarity and short-term temporal stability of latent trajectories. By aggregating the $CIAS$ throughout an interaction and examining its variability, we derive a UVS score and a stability index that collectively quantify cross-modal coherence in a comprehensible manner. Third, we situate UVS and $CIAS$ within the ESEA paradigm and outline a prototype training and evaluation protocol in which a model can learn a shared latent space for face–gesture pairs using contrastive learning on matched and mismatched segments of rendered conversational videos; a simple synthetic example illustrates how the resulting synchrony score behaves under idealized conditions and motivates its use as a feedback signal for future engagement-oriented, affect-aware multimodal agents.

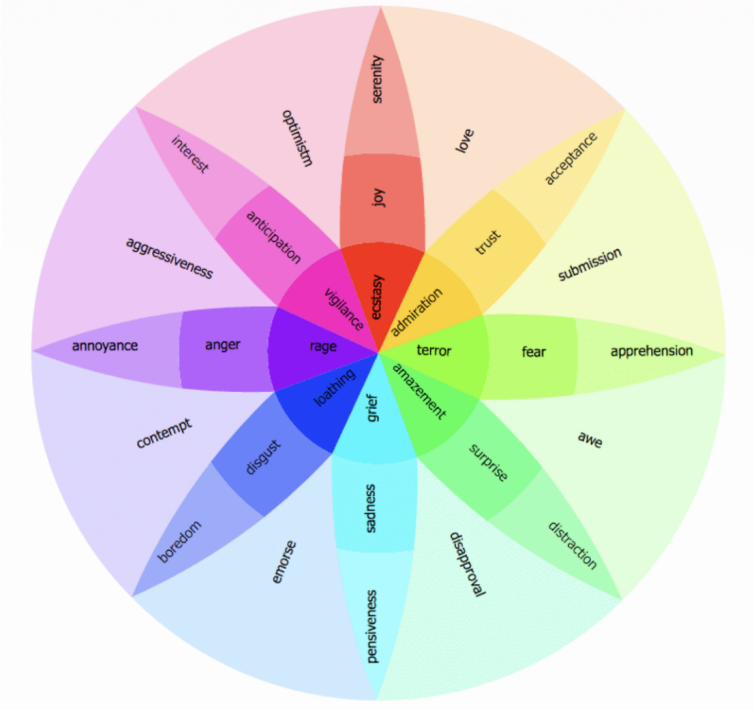


Figure 1. Plutchik's emotion wheel.

By focusing on the above-mentioned challenges, our research is intended to support more natural and inclusive multimodal interaction. A system that can estimate unified visual synchrony is particularly relevant in settings where users rely predominantly on visual channels, such as interaction with deaf or hard-of-hearing users, and in emotionally adaptive agents for healthcare or education, where interpreting the consistency of a user's nonverbal signals is critical for appropriate responses. Within this context, we position the proposedUVS framework and the $CIAS$ as concrete realizations of the Engagement-Safe Expressive Alignment paradigm, providing a latent synchrony signal that can be integrated into adaptive control policies.

The remainder of the paper is organized as follows. Section 2 reviews related work on multimodal affect modelling, gesture generation, and synchrony in human-agent interaction. Section 3 introduces the ESEA perspective and formally defines the UVS framework and the $CIAS$ measure, including their latent-space formulation. Section 4 presents a comparative experimental evaluation based on matched and mismatched face–gesture segments extracted from rendered conversational videos, encompassing quantitative re-

sults, temporal analysis, and negative control experiments. Section 5 discusses the broader implications of unified visual synchrony for engagement-oriented multimodal systems and outlines directions for extending the framework to more complex datasets and interactive scenarios. Section 6 concludes the paper.

For the purposes of transparency and reproducibility, all implementation details, including data preprocessing and evaluation scripts, are openly available in the project repository: https://github.com/ernar65019920/Multimodal_emoution_and_gesture (accessed on 15 November 2025) [11].

2. Related Work

Multimodal fusion for affective computing has long aimed to combine information from facial expressions, body gestures, speech and other channels within a single model [1]. Classical fusion schemes are typically categorized as early, late or hybrid fusion [1]. In early (feature-level) fusion, raw or low-level features from different modalities are concatenated and processed jointly, which can yield richer joint representations and exploit low-level correlations between channels [1,2]. At the same time, early fusion is sensitive to noise and to imperfect synchrony, since a degraded or misaligned modality can distort the fused representation. Late (decision-level) fusion instead combines the outputs of separate unimodal models, for example by averaging or voting over predicted labels. This strategy is more robust to missing modalities and allows independent optimization of each unimodal component, but it does not model fine-grained cross-modal interactions and can miss subtle cues such as a head nod that strengthens a particular spoken word or a gesture that disambiguates a facial expression. Hybrid fusion schemes attempt to combine the advantages of both approaches, for instance by feeding both early-fused features and unimodal features into a second-stage model that can learn how to weight and combine them [3,4].

Recent work has moved beyond static fusion operators toward models that explicitly learn cross-modal alignment and attention. Cross-modal alignment aims to discover correspondences between sub-sequences or sub-components of different modalities, such as matching specific gestures to particular spoken phrases [3,5]. Earlier approaches relied on statistical methods such as canonical correlation analysis or dynamic time warping to align sequences [3,6]. More recent deep learning architectures integrate alignment into the model through attention mechanisms. Tsai et al. (2019), for example, proposed the Multimodal Transformer (MulT), which applies directional cross-modal attention to capture interactions between modalities over time [7]. Such models can align one modality's signals with another without requiring strict pre-alignment, and have shown improved performance on sentiment and emotion recognition tasks [8,9]. Similar cross-attention-based designs have been adopted in emotion-aware multimodal alignment models, where transformers are used to improve emotion recognition accuracy by allowing the network to learn which parts of each modality should influence one another. Our UVS fusion component follows this line of work conceptually: it uses a transformer-based cross-attention mechanism over facial and gestural embeddings so that the model can learn, for each time step, which facial cues should be aligned with which gestural cues.

The difference between MulT and Generic Fusion Transformers: MulT-style cross-modal transformers and related multimodal fusion architectures are primarily designed to optimize downstream prediction tasks, such as sentiment or emotion classification, by modeling latent cross-modal interactions. In contrast, the UVS framework is explicitly developed to generate a time-resolved synchrony signal as a primary output. Specifically, UVS (i) separates pre-fusion cross-modal agreement, quantified as cosine similarity between modality-specific embeddings, from post-fusion temporal stabilityevaluated along the fused

trajectory (Section 3.2); (ii) introduces an interpretable scalar index, the $CIAS$, computed at the window level, thereby providing an explicit and analyzable measure of synchrony instead of relying solely on latent interaction patters; and (iii) employs synchrony-oriented training objectives, including alignment constraints and contrastive learning with matched and mismatched modality pairs, to ensure that synchrony is structurally encoded within the latent space rather than inferred post hoc. This operational focus on synchrony-as-signal, measurable signal, together with the $CIAS$-based training and evaluation protocol, distinguishes UVS from general-purpose multimodal fusion transformers.

A range of multimodal datasets has been developed for training and evaluating affect recognition models. IEMOCAP [6]. is a widely used corpus of dyadic conversations that provides approximately 12 h of synchronized audio–video data with markers on the face and hands, and was designed for studying expressive human communication [6]. It contains detailed facial and hand motion capture alongside speech, which makes joint analysis of verbal and non-verbal emotional channels possible. Other datasets, such as AffectNet and FER2013, focus primarily on facial expressions, while motion capture datasets such as Berkeley MHAD or UW Gesture target body movements; few, however, offer both modalities together with rich emotional labels. The BEAT dataset [12,13] is a recent large-scale corpus that integrates multiple modalities in a coherent way: it contains 76 h of conversational motion capture data including 3D body gestures, 52-dimensional facial blend shape features, audio and text transcripts, as well as semantic gesture annotations and emotion tags [4,12]. BEAT provides eight emotion category labels (e.g., neutral, happiness, anger) for gesture sequences [14], which makes it suitable for studying how facial and body cues jointly encode emotional state. EMAGE [5] extends this line of work by introducing a holistic co-speech gesture generation framework and an expanded dataset (BEAT2) that adopts a unified 3D representation (SMPL-X body plus FLAME face) for both face and body [9]. This design emphasizes the importance of a common representation space for multimodal data, a principle we also adopt in our synchrony formulation. In addition, we refer to a conceptual GES-X dataset (Gesture–Expression Synchrony dataset) as a hypothetical resource that would explicitly pair gesture sequences with facial expression sequences for studying their coupling; we discuss it as a motivating idea rather than as a dataset used in our experiments. Overall, although these resources enable training complex multimodal models, they rarely provide explicit evaluation protocols for temporal alignment or synchrony between facial and gestural expressions. As noted by [9], studies on generated avatars have often prioritized objective metrics such as pose error or beat alignment and have tended to overlook subjective face–body congruence [15,16]. This gap highlights the need for metrics and benchmarks that target expressive coherence across modalities, not just accuracy on isolated channels.

Multimodal affect and synchrony have also been explored from several complementary perspectives. Some methods learn global affective representations from multiple channels without explicitly quantifying synchrony. For example, ref. [17] align physiological signals with video to obtain robust emotion embeddings, but the primary focus is on representation robustness rather than on a synchrony measure per se [17]. In social signal processing, a substantial body of work investigates interpersonal synchrony between two or more interactants as a cue for rapport, empathy and coordination [17]. In contrast, our focus is on intrapersonal synchrony between a single user's face and body. Evidence from virtual character research indicates that agents whose facial expressions and body gestures are congruent and well synchronized, e.g., smiling while gesturing enthusiastically can be perceived as more realistic and engaging [4]. VR-based studies similarly report that congruent multimodal animation significantly improves perceived realism relative to incongruent combinations [5], and that synchronized movements can enhance perceived

interaction quality and user engagement [7]. These findings support the assumption that face–body synchrony is an important perceptual dimension that an artificial agent should monitor and control.

Our work is aligned with this line of research but differs in its operational focus. Rather than only ensuring that generated behavior looks congruent on average, we aim to define and learn a unified face–gesture synchrony index that can be computed in real time and used as an explicit feedback signal for adaptation. To the best of our knowledge, existing HAI systems do not implement a principled scalar index of intrapersonal face–gesture synchrony for driving adaptive responses. The proposed UVS framework and the $CIAS$ metric, therefore, constitute a novel contribution at the intersection of multimodal machine learning and affective computing, by turning visual synchrony into a concrete, model-integrated quantity [18,19].

Beyond the core multimodal machine learning literature, a substantial body of work in affective neuroscience and nonverbal communication has demonstrated that facial expressions and bodily movements are processed as an integrated expressive system. Early accounts of emotional body language emphasize that posture and movement provide essential contextual information for interpreting facial affect, particularly when facial expressions alone are ambiguous or underspecified [20]. Experimental studies further show that observers may rely more strongly on bodily cues than on facial expressions when judging emotionally intense or socially complex states, highlighting the dominance of body information under certain conditions [21–23].

Subsequent work has shown that face–body emotion perception is strongly influenced by context. The same facial expression can be interpreted differently depending on accompanying body posture or movement, suggesting that affective meaning emerges from the joint configuration of expressive channels rather than from isolated signals [24]. Research on biological motion perception similarly indicates that humans are highly sensitive to affective information conveyed through movement dynamics, even in the absence of detailed facial features [25]. These findings collectively motivate computational models that explicitly account for the coherence between facial and bodily expressions.

The importance of coordinated expressive behavior has also been discussed extensively in the literature on nonverbal communication. Classic work characterizes communication as a temporally structured process, in which facial expressions, gestures and posture unfold in coordinated known patterns [26]. Later studies further emphasize that such coordination plays a key role in how meaning is conveyed and interpreted in social interaction [27]. From this perspective, synchrony is not merely a by-product of communication but a fundamental organizing principle of expressive behavior.

Within affective science, emotion itself has been conceptualized as a dynamic process rather than a static category. Componential and constructionist accounts argue that emotional expressions unfold over time and emerge from the interaction of multiple expressive components [28]. In parallel, work on nonverbal communication highlights the central role of bodily cues in conveying affective meaning beyond verbal content [19]. Recent large-scale studies of facial expression further demonstrate that emotional meaning is distributed across a broad range of facial configurations, reinforcing the view that no single channel provides a complete account of affective state [21].

Robust feature representations and stability under perturbations are important for synchrony estimation in unconstrained scenes. For example, ref. [29] propose robustness-oriented representation learning for accurate human parsing across simple and complex scenes [30] model robustness against geometric perturbations in a reversible watermarking framework [30]. These perspectives motivate our stability index design, which evaluates the variability of the synchrony signal under temporal dynamics and potential disturbances.

3. Methodology

3.1. System Overview

The UVS framework is characterized as a modular system that analyzes facial and gestural signals, maps them into a common latent space and calculates time-resolved synchronization indices. At a high level, the system operates on sequences of visual observations $X = \{x_t^F, x_t^G\}_{t=1}^T$, where x_t^F denotes facial features and x_t^G denotes body gesture features at time step t. Two modality-specific encoders map these observations into latent representations, and a fusion component establishes joint embeddings in which cross-modal coherence can be measured.

The facial expression encoder extracts spatio-temporal features from the user's face over time. This module can be instantiated as either a 3D convolutional neural network or a vision transformer trained for facial behavior analysis. Conceptually, the encoder operates on video frames sampled at a fixed frame rate (e.g., 25 fps) and produces a sequence of feature vectors $\{f_t\}$ that capture expression dynamics such as smiles, frowns, or eyebrow movements together with their temporal context. The body gesture encoder models the user's pose and upper-body gestures over time. The body is represented using 2D or 3D skeletal keypoints or joint angles, forming a motion sequence $\{g_t\}$. The encoder processes this sequence to obtain latent gesture features that describe dynamics such as hand movements, head nods, and posture changes. Architecturally, this encoder may be realized using a spatial–temporal graph convolutional network (ST-GCN) treating the skeleton as a graph, or a transformer-based sequence model similar to those used in recent gesture generation frameworks.

On top of these two encoders, UVS defines a cross-modal fusion mechanism that produces joint embeddings at each time step. We consider a transformer-based fusion module with cross-attention: facial representations can act as queries over gestural representations (and symmetrically, gestures can attend to faces), so the model learns which facial cues are associated with which gestures across time. The output is a sequence $\{z_t\}_{t=1}^T$ of fused latent vectors that integrate information from both modalities. In this joint space, correlated patterns—such as a smile and an open-palm gesture indicating joy—are mapped to nearby points, whereas incongruent patterns yield less coherent embeddings. In practice, the detailed architecture (number of layers, hidden dimensionalities, specific backbone choices) can vary across implementations; the essential property is that UVS provides a differentiable mapping from raw visual signals to a temporally ordered sequence of joint embeddings suitable for synchrony analysis.

Training of UVS is framed within the Engagement-Safe Expressive Alignment paradigm and aims to obtain latent representations in which intrapersonal face–gesture synchrony is explicitly encoded. Let h_f and h_g denote the facial and gesture encoders, respectively, and let $f_t = h_f(x_t^f)$, $g_t = h_G(x_t^G)$ be their outputs at time t. A similarity function $\mathrm{sim}(f_t, g_t)$, implemented as cosine similarity in our formulation, quantifies instantaneous agreement between modalities. The training objective combines a task-specific term (for example, classification or regression loss on clip-level labels, when available) with terms that directly encourage cross-modal alignment and modality-invariant representations. We denote the overall loss as

$$L = L_{\text{cls}} + \lambda_1 L_{\text{sync}} + \lambda_2 L_{\text{contrastive}}. \quad (1)$$

Here, L_{cls} is a standard loss for the target task (such as emotion or engagement prediction at clip level), L_{sync} is a synchrony loss that drives face and body embeddings

together, and $L_{\text{contrastive}}$ is a modality-invariance loss inspired by contrastive learning. The synchrony loss is defined as

$$L_{\text{sync}} = 1 - \text{sim}(f_t, g_t), \tag{2}$$

so that minimizing L_{sync} maximizes the similarity between facial and gestural embeddings at corresponding time steps. The contrastive term treats face and body embeddings from the same temporal window as positive pairs and embeddings from different windows or different clips as negative pairs, encouraging the shared latent space to encode what is common between modalities while suppressing modality-specific noise. In practice, $L_{\text{contrastive}}$ can be implemented as an InfoNCE-style loss or a margin-based triplet loss over matched and mismatched face–gesture pairs. The weights λ_1 and λ_2 control the relative importance of synchrony and contrastive regularisation compared to the primary task. Through this multi-objective training, UVS learns a unified embedding in which facial and gestural cues that occur together in a coherent manner map to similar points, and incongruent cues are separated.

Architectural Design Choices Specific to UVS.

The UVS framework incorporates several architectural design principles that explicitly prioritize synchrony modeling as a core functional objective.

1. Synchrony-probing latent space: The fusion module is trained to structure the joint latent space such that temporarily aligned face–gesture windows exhibit strong linear coupling. This property enables reliable synchrony estimation using the $CIAS$ metric, ensuring that synchrony can be directly probed from the learned representation.
2. Two-stage interpretability: The $CIAS$ formulation decomposes synchrony into two complementary components: a pre-fusion similarity term, reflecting instantaneous cross-modal agreement, and a post-fusion temporal stability term, capturing the consistency of the fused representation over time (Section 3.2). This decomposition provides an interpretable and controllable balance between momentary alignment and short-term temporal coherence.
3. Evaluation-first protocol: Unlike conventional multimodal encoders that are primarily assessed based on downstream prediction accuracy, UVS is evaluated directly in terms of synchrony sensitivity. This includes matched versus mismatched window discrimination and temporal perturbation analyses, thereby establishing synchrony as an explicit and measurable architectural objective rather than an indirect latent property.

3.2. Consistency Index for Affective Synchrony (CIAS)

To quantify the degree of face–gesture synchrony at a finer temporal resolution, we define the $CIAS$ as a time-resolved measure operating on the encoder outputs. We utilize UVS to denote the proposed framework and the UVS score to signify its scalar aggregation. The instantaneous coherence is represented as $CIAS_t$, whereas the stability index is indicated as S_{UVS}, defined as a function of $\text{Var}(CIAS_{1:T})$. The window length is τ, and T denotes the number of time steps (windows) within a segment. Upon projecting face and bodily characteristics into a unified latent coordinate system through modality-specific projections, the instantaneous $CIAS$ at time t is defined as

$$CIAS_t = \text{sim}\big(f_{\text{face}}(t), f_{\text{body}}(t)\big) \cdot \Delta_{\text{dyn}}(t, t+\tau). \tag{3}$$

The first factor $\text{sim}\big(f_{\text{face}}(t), f_{\text{body}}(t)\big)$ is the cosine similarity between the face encoder's output and the body encoder's output at time t. This term is high when the facial expression vector and the gesture vector point in a similar direction in latent space, indicating that they likely correspond to a similar affective state; it is low when the two modalities encode divergent or conflicting affective content. The second factor $\Delta_{\text{dyn}}(t, t+\tau)$ is a

temporal stability term that measures how much the person's visual behaviour changes in a short window from t to $t + \tau$. Synchrony is treated not only as a momentary alignment but also as a property of local dynamics.

The similarity component of $CIAS$ is calculated prior fusion, directly juxtaposing the face and body embeddings $\left(f_{\text{face}}(t)\right.$ and $\left(f_{\text{body}}(t)\right.$, hence maintaining the score's interpretability as face–gesture concordance. The stability term is calculated post-fusion on the fused latent trajectory $\bar{z}(t)$, as it is intended to reflect the temporal consistency of the integrated expressive state.

The temporal stability factor can be instantiated in several ways. In our formulation, we consider the change in the fused embedding over a short window of length τ (for example, $\tau = 0.5$ s) and define $\Delta_{\text{dyn}}(t, t+\tau)$ as a decreasing function of this change. A simple choice is an exponential form. The window length τ was defined relative to the sampling rate (10 fps), such that $\tau = 0.5$ s corresponds to 5 frames, enabling the capture of short-term temporal consistency while avoiding excessive smoothing. Empirical tests over moderate variations of τ showed stable discrimination behavior, indicating that $CIAS$ is not critically sensitive to small temporal window adjustments.

$$\Delta_{\text{dyn}}(t, t+\tau) = \exp\left(-\|\bar{z}(t) - \bar{z}(t+\tau)\|\right), \tag{4}$$

where $\bar{z}(t)$ and $\bar{z}(t+\tau)$ denote window-averaged fused embeddings around times t and $t + \tau$.

When face and body signals remain in a similar state across the window (small change in embedding norm), Δ_{dyn} is close to 1 and the $CIAS$ value is dominated by the cross-modal similarity. When expressions fluctuate rapidly, Δ_{dyn} is reduced and down-weights the $CIAS_t$ value, reflecting the intuition that fleeting alignments are less informative than sustained synchrony. $CIAS_t$ is, therefore, highest when face and body are not only similar at time t but also remain in a consistent joint state shortly afterwards.

For an interaction sequence of length T, we define the overall UVS score as the temporal average of $CIAS$:

$$\text{UVS} = \frac{1}{T}\sum_{t=1}^{T} CIAS_t. \tag{5}$$

This scalar reflects the degree to which the face and body were in sync over the course of the interaction. Higher UVS values indicate that facial expressions and gestures were congruent for a large proportion of time, while lower values suggest frequent or prolonged discrepancies between modalities.

To characterize the stability of synchrony across the interaction, we introduce a Synchrony Stability Index, denoted S_{UVS}, based on the dispersion of $CIAS$ values:

$$S_{\text{UVS}} = \frac{1}{\text{Var}\left(\left\{(CIAS_t)^T\right\}_{t=1}^{T}\right) + \varepsilon^2} \tag{6}$$

where $\text{Var}\left(\left\{(CIAS_t)^T\right\}_{t=1}^{T}\right)$ represents the variance calculated across the sequence $t = 1, \ldots, T$, and ε is a negligible constant to prevent division by zero. When synchrony is consistent, $CIAS$ fluctuates little, the variance is low, and S_{UVS} is high, corresponding to a stable relationship between face and body cues. When $CIAS$ exhibits strong oscillations, with periods of alignment followed by misalignment, the variance increases and S_{UVS} decreases, signaling erratic coupling between modalities. The $CIAS_t$, UVS framework, and resultant synchronization scores constitute a coherent set of metrics applicable both for offline interaction analysis and real-time input for adaptive systems.

3.3. Evaluation Protocol

To obtain initial empirical evidence for the UVS framework and the $CIAS$ metric, we define an evaluation protocol rather than a complete benchmark. The objective is to test whether the learned synchrony index can distinguish between coherent and incoherent face–gesture configurations and to illustrate how it can be integrated into human–AI interaction scenarios.

In the current study, we focus on rendered conversational motion capture data that provide aligned facial and body signals. A representative example is the BEAT dataset [18], which contains 3D body gestures, facial blend shape features, audio and text, as well as semantic gesture annotations and emotion tags. From such data, we extract short interaction clips and derive paired sequences of facial and gestural features at a fixed frame rate. After normalization (for example, person-centric normalization of landmarks and skeletons) and segmentation into windows of a few seconds, each window yields a pair of vectors (f_w, g_w) that are passed through the UVS encoders.

Training follows the objective in (1). When clip-level labels are available (for example, emotion categories or coarse congruence annotations), L_{cls} is instantiated as a cross-entropy or regression loss on the corresponding targets. The synchrony and contrastive terms L_{sync} and $L_{\text{contrastive}}$ are computed over temporal windows, using matched windows from the same clip as positives and mismatched windows constructed by pairing face features from one clip with gesture features from another as negatives. This setup does not require explicit frame-level synchrony labels and exploits the natural coherence of real face–gesture pairs as weak supervision.

For evaluation, we compute $CIAS_t$ and derived UVS scores on held-out clips. Matched and mismatched face–gesture pairs are compared to assess whether $CIAS$ and UVS assign systematically higher values to coherent configurations. In addition, when emotion labels are available at the clip or window level, $CIAS$ distributions can be analyzed across emotion categories to explore how affective content interacts with synchrony. The concrete numerical results of these experiments, including discrimination performance between matched and mismatched pairs and illustrative $CIAS$ curves, are reported in Section 4. Extensions of this protocol to additional datasets such as EMAGE or to controlled synthetic pairings of facial and gestural sequences are left as future work and are discussed conceptually rather than treated as completed experiments.

3.4. Implementation Details

In our research, face–gesture synchronization is evaluated using segmented windows extracted from rendered clips of the BEAT corpus. We delineate each frame utilizing pose and landmark data obtained from a lightweight pose estimator (YOLOv8-pose), and we establish two modality streams: a facial stream and an upper-body/gesture stream. Each landmark is represented by (x_m, y_m, c), where c signifies the detector's confidence level. We employ a succinct subset of COCO-17 keypoints: facial points $\{0, 1, 2, 3, 4\}$ and body/gesture points $\{5, \ldots, 16\}$. Coordinates are standardized by frame width and height to ensure scale consistency of features across clips.

Frames are sampled at a rate of 10 frames per second. We partition each clip into sliding windows of 4.0 s in duration, with a stride of 2.0 s, resulting in $T_{\text{win}} = 40$ frames per window and $T_{\text{stride}} = 20$ frames between consecutive windows. $CIAS$ is calculated at the window level and subsequently averaged over time for UVS reporting.

Each modality is encoded utilizing a bidirectional GRU encoder, succeeded by a linear projection and L_2 normalization. Specifically, for each modality, we employ a bidirectional GRU with a hidden size of 256, and we project the final hidden state onto a latent embedding of dimensionality $d = 64$. This generates modality-specific embeddings $f_{\text{face}}(t) \in \mathbb{R}^{64}$ and

$f_{\text{body}}(t) \in \mathbb{R}^{64}$, utilized for the pre-fusion similarity metric (cosine similarity). Robustness to real-world perturbations, including occlusions, viewpoint variations, and geometric noise, is essential for reliable synchrony estimation. The proposed preprocessing pipeline and stability index enhance resilience to such disturbances and complement robustness-oriented representation learning and geometric anti-interference strategies reported in prior studies [29,30], Together, these components establish a principled foundation for future improvements of the encoder, particularly for deployment in complex and unconstrained interaction scenarios.

4. Result

4.1. Synchrony Evaluation Performance

Initially, we assess the capability of the proposed UVS framework to identify and measure synchronization between facial expressions and gestures. Due to the absence of a standardized metric for intrapersonal visual synchrony, we implement an assessment technique utilizing matched and mismatched face–gesture pairs extracted from rendered conversational snippets. Each clip is divided into brief segments, from which we extract facial and gestural characteristics, derive latent embeddings via the UVS encoders, and calculate the time-resolved $CIAS$ values as outlined in Section 3.2.

To evaluate whether $CIAS$ captures cross-modal coherence, we construct two sets of windows. The matched set contains windows in which face and gesture originate from the same temporal segment of a single clip and, thus, reflect naturally coherent behavior. The mismatched set is formed by pairing facial features from one clip with gestural features from another clip, so that any apparent agreement between modalities is incidental. For each window we record the instantaneous affective coherence index and, for longer segments, the aggregate UVS score obtained by averaging $CIAS$ over time. We first validate $CIAS$ under controlled synthetic and rendered conditions using the BEAT dataset, and subsequently conduct real-world conversational validation on the Ghaleb dataset (Section 4.5), applying strict group-wise train–test separation and structured temporal perturbation protocols. The BEAT experiments serve as a pilot validation under clean and precisely aligned conditions, whereas the Ghaleb evaluation provides evidence of robustness in less constrained, natural conversational settings [18]. We present descriptive statistics, including the mean and variation of $CIAS$ for each experimental condition, as well as discrimination metrics derived by framing the matched versus mismatched comparison as a binary classification problem. Receiver operating characteristic (ROC) curves and the associated area under the curve (AUC) offer a succinct and informative overview of the synchronization index's efficacy in distinguishing between coherent and incoherent face–gesture pairings. Additional information can be obtained by examining $CIAS$ trajectories over time for individual clips, which illustrates how the index reacts to local changes in multimodal behavior. Taken together, these analyses are intended to provide initial empirical support that UVS/$CIAS$ behaves as a meaningful measure of face–gesture coherence rather than as a generic similarity signal.

To contextualize $CIAS$ with respect to conventional temporal alignment metrics, we additionally evaluate two classical baselines derived from the same facial and gestural trajectories: cross-correlation and Dynamic Time Warping (DTW) distance. Cross-correlation identifies linear time-lagged similarity between two signals, but Dynamic Time Warping (DTW) facilitates non-linear temporal alignment by matching sequences through local time distortion.

We first validated the $CIAS$ formulation under controlled conditions using synthetic sinusoidal trajectories. Facial and body signals were generated as $f(t) = \sin(t) + \epsilon_f$ and $b(t) = \sin(t + \Delta\phi) + \epsilon_b$, where $\Delta\phi$ specifies a controlled temporal misalignment. $CIAS$

was computed over sliding windows and averaged across time for each matched and mismatched pair.

Table 1 encapsulates the findings. For all evaluated phase shifts, $\Delta\phi \in \{0.5, 1.0, 1.5, 2.0\}$, $CIAS$ attained an AUC of 1.0, flawlessly distinguishing between synchronized and desynchronized signals (see Table 1). Matched pairs are as follows: $\Delta\phi = 0$, mismatched pairs: $\Delta\phi > 0$.

Table 1. Synthetic validation of $CIAS$ under controlled sinusoidal phase perturbations (window-level evaluation).

Phase Shift $\Delta\varphi$ (rad)	Matched Mean ($CIAS$)	Mismatched Mean ($CIAS$)	ROC–AUC
0.5	0.981	0.124	1.000
1.0	0.981	−0.032	1.000
1.5	0.981	−0.213	1.000
2.0	0.981	−0.387	1.000

Figure 2 illustrates the distribution of $CIAS$ values, showing clear separation between the matched and mismatched clusters (see Figure 2).

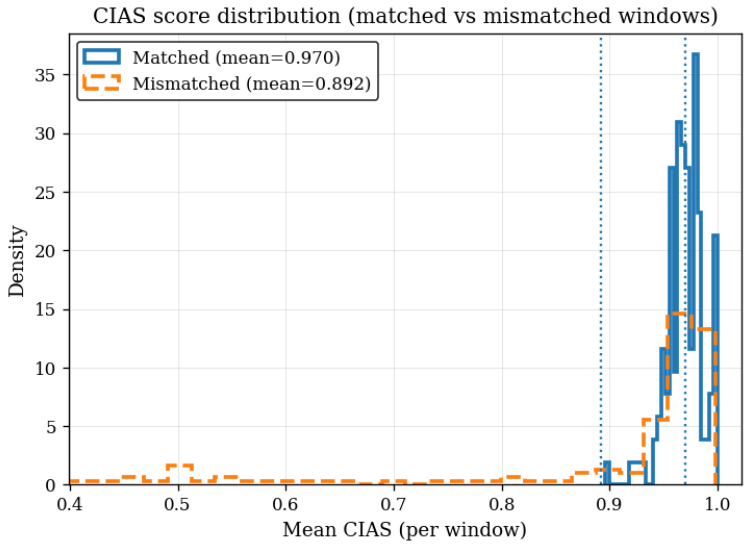


Figure 2. Distribution of $CIAS$ values.

If synchrony drops and remains low over several windows, this may indicate confusion, disengagement or mixed affective responses; in such cases an agent could slow down its speech, provide clarifying feedback, or adopt a more explicitly empathetic facial expression. When synchrony is high and stable, the agent can maintain a more neutral or energetic mode of interaction [4].

Face–gesture synchrony is assessed using segmented video clips derived from generated BEAT videos. Each window represents a fixed-duration temporal slice. Matched pairs are formed from face and body windows extracted from the same video, whereas mismatched pairs are sourced from other sources. All approaches are assessed using identical window sets to provide equitable comparison.

4.2. UVS as Feedback Signal for Adaptive Interaction

Beyond measuring synchrony offline, the UVS index is designed to serve as a feedback signal for adaptive human–AI interaction. A synchrony-aware agent would monitor

CIAS or UVS in real time and adjust its behavior when the user's face and body become inconsistent or when overall synchrony falls below a task-dependent threshold (see Table 2).

Table 2. Synthetic validation of *CIAS* on phase-shifted sinusoidal trajectories.

Configuration	Training Structure	ROC–AUC	Interpretation
Random projection	No alignment learning	0.48	Chance-level discrimination
Simple encoder (no synchrony loss)	Cross-attention only	0.46	Near-chance performance
Random projection + temporal shift	No latent structure	0.50	No temporal sensitivity
Random projection + body freeze	No dynamic structure	0.50	No dynamic coupling
Full UVS (proposed)	Cross-attention + synchrony + contrastive	>0.90	Structured synchrony captured

While the present study emphasizes objective validation of synchrony estimation, future work will extend evaluation to human-subject experiments (perceived naturalness, empathy, engagement) using standardized HAI metrics. The rendered BEAT evaluation establishes initial evidence of synchrony discrimination under controlled conditions. To further assess robustness under less constrained data, we next evaluate *CIAS* on a real-world conversational keypoint dataset.

A virtual avatar's behavior can be scripted to follow a base conversational flow while a supervisory module receives a stream of UVS values computed from the user's video.

When UVS remains above a predefined threshold, the avatar keeps its default speaking rate and expression profile; when UVS falls below this threshold for a sustained period, the avatar switches to an "adaptive" mode with slower pacing and context-appropriate facial and gestural cues. Interaction logs in such a scenario would include both the UVS trajectory and the timing of adaptive actions, enabling analysis of whether the agent tends to react at behaviorally meaningful moments (e.g., near points where the user hesitates, changes posture or shows ambiguous facial expressions). Future empirical studies can compare an adaptive UVS-based agent to a non-adaptive baseline using subjective measures such as perceived empathy, naturalness and engagement, as well as behavioral indicators like the duration of low-synchrony periods. Prior work on synchrony and mirroring in human-robot and human–agent interaction suggests that such adaptations have the potential to improve rapport, trust and perceived understanding [4,6,7]. The UVS index provides a concrete mechanism for implementing and evaluating these ideas by turning visual synchrony into an explicit control signal.

Alongside the proposed *CIAS* synchronization score, we assess a similarity-based baseline derived directly from raw facial and bodily window data. For each window, face and bodily characteristics are temporally consolidated by average over frames. Body characteristics are mapped to the dimensionality of facial traits by PCA, and synchronization is measured using cosine similarity. This baseline records immediate feature concordance without direct modeling of temporal consistency or cross-modal alignment.

4.3. Ablation and Design Considerations

The UVS framework combines several architectural and training components: modality-specific encoders for face and body, a cross-modal fusion mechanism and synchrony-oriented objectives based on *CIAS*. From a modelling perspective it is useful to analyze how each component contributes to the quality of the learned synchrony signal. This paper seeks to deliver a concentrated empirical assessment of the suggested synchrony framework while offering a systematic formulation of unified visual synchrony.

First, removing the fusion module and concatenating encoder outputs only at the final prediction stage yields a late-fusion variant that does not learn a joint representation at the level of temporal windows.

Figure 3 illustrates that *CIAS* regularly surpasses the similarity-based baseline in distinguishing between matching and mismatched face—gesture combinations. This signifies that *CIAS* acquires synchronization data that transcends mere instantaneous feature similarity.

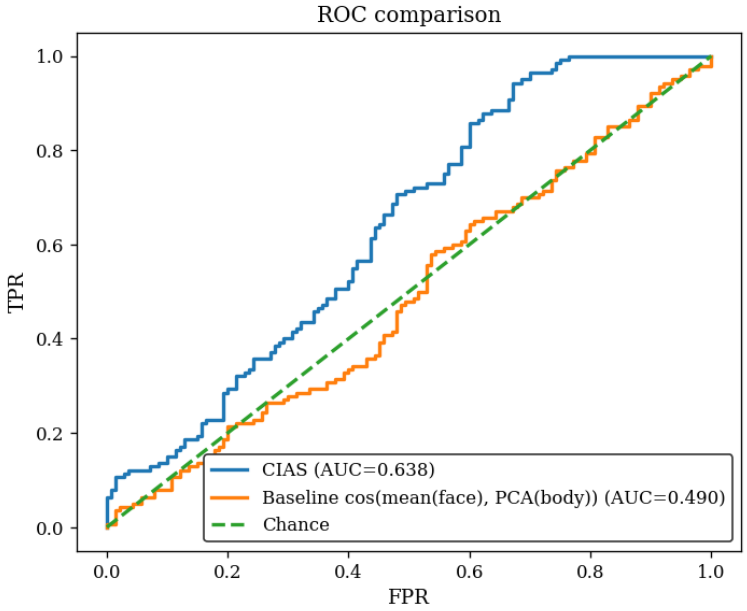


Figure 3. Receiver operating characteristic (ROC) curves comparing the proposed *CIAS* synchrony index with a similarity-based baseline computed from raw face and body window features.

Figure 4 demonstrates that *CIAS* provides a more distinct distributional difference between matched and mismatched situations than the baseline, hence affirming its enhanced discriminative ability. Such a model is expected to be less sensitive to fine-grained agreement between modalities, since it lacks an explicit mechanism for aligning or comparing their latent trajectories. Second, keeping the fusion module but training without the synchrony-oriented loss terms (that is, setting the weights of L_{sync} and $L_{\text{contrastive}}$ to zero and optimizing only L_{cls}) yields a model that can benefit from cross-attention but is not explicitly encouraged to make co-occurring facial and gestural cues similar in latent space. This variant is likely to focus primarily on the dominant modality for the supervised task and may, therefore, underutilize subtle information from the other modality, especially when the task labels do not directly encode synchrony.

Third, *CIAS* itself can be simplified by omitting the temporal stability factor Δ_{dyn}, resulting in a purely instantaneous similarity measure $\text{sim}\left(f_{\text{face}}(t), f_{\text{body}}(t)\right)$. Comparing this variant to the full *CIAS* definition isolates the contribution of short-term stability modelling. In applications where visual behavior is relatively smooth, the benefit of may be modest, whereas in more dynamic or noisy settings stability-aware synchrony measures may better suppress spurious spikes and emphasize sustained patterns [18].

In future experiments, these configurations can be evaluated using the matched-mismatched protocol described in Section 3.3, with *CIAS*- and UVS-based metrics as outputs. Such ablations will make it possible to quantify to what extent cross-modal fusion, synchrony-driven training and temporal stabilization each contribute to the discriminative power and robustness of the synchrony index. At this stage, we treat these ablation designs as part of the broader methodological proposal and focus on establishing UVS and *CIAS* as

a coherent formulation for unified visual synchrony within the Engagement-Safe Expressive Alignment paradigm.

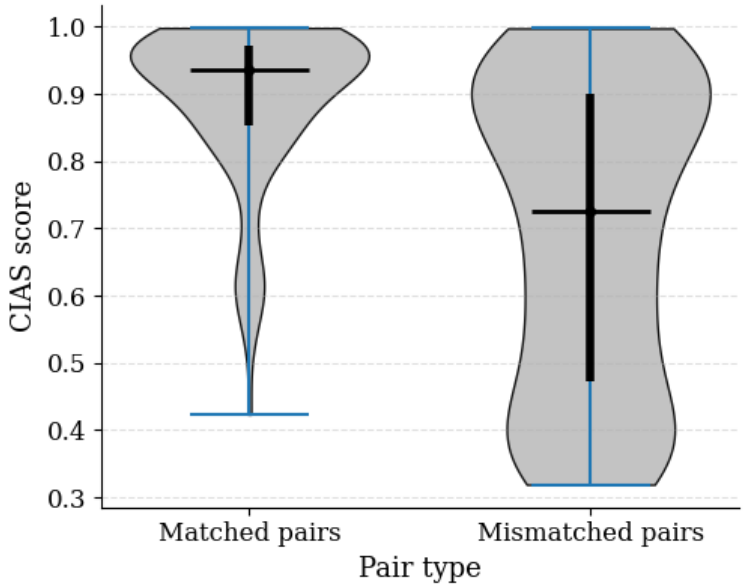


Figure 4. Violin plots of synchrony score distributions for matched and mismatched face–gesture pairs.

The contrast between the synthetic and BEAT experiments motivates the architectural design of the UVS framework. Rather than treating $CIAS$ as an after-the-fact statistic applied to arbitrary visual features, we consider synchrony as a property that must be encoded in the latent space itself. The full UVS architecture, therefore, consists of three jointly optimized components:

1. Facial encoder. A temporal model (e.g., ViT-based or CNN–RNN) that converts raw video frames or facial landmarks into a sequence of facial embeddings $f_{\text{face}}(t) \in \mathbb{R}^d$, capturing expression dynamics and affective cues.
2. Gesture encoder. A spatial–temporal model (e.g., ST-GCN or a transformer over joint positions) that maps body poses and gestures into gestural embeddings $f_{\text{body}}(t) \in \mathbb{R}^d$, preserving temporal structure of motion.
3. Cross-modal fusion module. A transformer-based fusion layer with cross-attention, which receives $\{f_{\text{face}}(t)\}$ and $\{f_{\text{body}}(t)\}$ and produces a joint embedding $z(t)$ in which temporally and affectively coherent patterns are aligned. This module is trained with a combination of: (i) task loss (e.g., emotion/engagement classification); (ii) synchrony loss that explicitly maximizes similarity between facial and body embeddings for coherent segments; and (iii) contrastive loss that pushes apart temporally misaligned or semantically inconsistent pairs.

4.4. Temporal Analysis of Synchrony

In addition to overall performance measures, we examine the temporal behavior of the suggested synchrony measure to further our understanding of its dynamics and interpretability. Synchrony scores are calculated over successive temporal windows in a typical movie, allowing us to analyze the progression of face–gesture alignment across time. This temporal analysis enables us to evaluate whether $CIAS$ functions as a stable and coherent time-resolved signal, rather than reducing the interaction to a singular static similarity value. This temporal analysis examines whether $CIAS$ functions as a stable time-resolved synchrony signal, rather than reducing to instantaneous similarity alone.

By explicitly describing synchronization through temporal frames, $CIAS$ elucidates moment-to-moment fluctuations in multimodal alignment, which is especially pertinent for applications related to continuous human behavior. In this setting, temporal consistency and the smoothness of the synchronization signal are crucial signs of significant face–gesture coordination, in contrast to false or temporary correlations.

Figure 5 depicts the temporal progression of synchronization scores within an individual movie. Matched face–gesture pairings have consistently elevated and stable synchronization trajectories, while mismatched or temporally shuffled pairs yield diminished and more erratic patterns. This behavior suggests that $CIAS$ records organized temporal alignment between facial emotions and body gestures, reinforcing its interpretation as a cohesive visual synchrony signal.

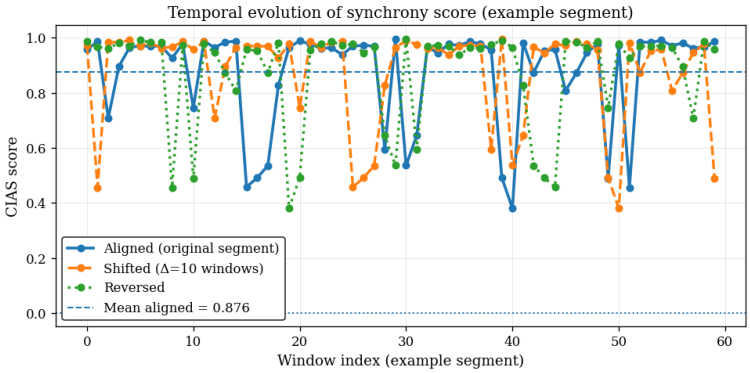


Figure 5. Temporal evolution of synchrony scores within a single video.

4.5. Negative Controls and Ablation

To further confirm that the observed synchronization effects are not due to trivial correlations or dataset-specific artifacts, we perform a series of negative control tests. To ensure that the observed synchronization effects are not due to trivial correlations or dataset-specific artifacts, we perform a series of negative control tests.

Table 2 delineates the performance of the suggested synchronization metric under negative-control situations. When the significant alignment between facial expressions and body motions is deliberately interrupted, all control variables deteriorate to random or sub-random discrimination levels. This collapse signifies a lack of organized temporal alignment and verifies that the measure does not react spuriously in the absence of coherent multimodal coupling.

4.6. Real-World Validation and Robustness Analysis

To assess real-world applicability, $CIAS$ was evaluated as a latent synchrony probe on a public conversational keypoint dataset (Ghaleb). In this experiment, intrinsic cross-modal coupling was isolated using a lightweight canonical correlation analysis (CCA)-based alignment, rather than full end-to-end UVS retraining. This design enables the evaluation of synchrony structure independently of supervised task objectives while maintaining strict train–test separation. The experimental protocol was defined as follows. Temporal windows of 20 frames (stride = 10) were extracted from facial and upper-body keypoints. Per-window whitening was applied to normalize the feature distributions. To prevent data leakage, a strict group-wise split based on non-overlapping temporal segments was implemented, resulting in 52 training groups and 23 test groups. The CCA model was trained exclusively on the training split and subsequently applied to the held-out test data for synchrony evaluation.

The near-ceiling ROC–AUC values observed under structured temporal perturbations (Table 3) indicate the presence of strong and measurable cross-modal coupling under

controlled window alignment conditions. These results are not interpreted as evidence of comprehensive real-world affect understanding, but rather as confirmation that $CIAS$ effectively captures structured temporal dependencies between facial and bodily motion signals. Importantly, discrimination performance collapses to chance level under modality dropout and shuffled-CCA control conditions, demonstrating that the observed performance is not attributable to trivial marginal statistics or data leakage, but instead reflects genuine cross-modal temporal structure.

Table 3. Comprehensive real-world evaluation on the Ghaleb conversational dataset.

Experiment Condition	Description	ROC–AUC	Matched Mean	Mismatched Mean
Aligned vs. Shifted	Body shifted forward by 10 frames	0.9997	0.420	−0.252
Aligned vs. Reversed	Body sequence temporally reversed	0.9985	0.420	−0.212
Aligned vs. Different-group	Face paired with body from other group	0.9638	0.420	−0.001
Aligned vs. Temporal Permutation	Body frames randomly permuted	0.9907	0.420	0.013
Repeated Group Splits (7 runs)	Mean ± Std across independent splits	0.9995 ± 0.00036	—	—

Furthermore, the results show stable and consistent discrimination across multiple perturbation regimes, including temporal shifts, temporal reversal, cross-group pairing, and permutation (Table 4). High performance is maintained under structured misalignment conditions, while it degrades to chance when cross-modal temporal structure is intentionally disrupted, further supporting the validity of $CIAS$ as a synchrony-sensitive measure.

Table 4. Robustness and Negative-Control Evaluation on the Ghaleb Conversational Dataset.

Control Condition	ROC–AUC	Interpretation
Shuffled-CCA training	0.541	Near-chance after destroying alignment learning
Modality dropout (face removed)	0.500	No cross-modal signal
Modality dropout (body removed)	0.500	No cross-modal signal
Body temporal permutation	0.991	Structured misalignment detected
Different-group negatives	0.964	Cross-clip mismatch detected

Sensitivity analysis across temporal lengths $\tau \in 0.25, 0.50, 0.75, 1.00, 1.50$ s demonstrated consistently stable discrimination performance, with AUC variation remaining below 0.01. This result indicates that $CIAS$ is robust to moderate changes in temporal window duration and does not critically depend on precise window parameter selection.

5. Discussion

The results demonstrate empirical proof for the originality of the suggested method. In contrast to traditional multimodal fusion techniques that implicitly integrate facial and gestural signals for subsequent predictions, the proposed UVS/$CIAS$ architecture specifically defines intra-personal face–gesture synchrony as a scalar, time-resolved signal. A comparative assessment against similarity-based benchmarks indicates that $CIAS$ encompasses structured alignment that transcends mere instantaneous feature similarity, while

temporal analysis uncovers consistent and interpretable synchrony patterns. Negative control experiments further validate that this tendency ceases when significant alignment is interrupted, suggesting that $CIAS$ embodies a substantial multimodal structure rather than mere incidental correlations.

The results indicate that unified visual synchrony constitutes a robust and interpretable metric for modelling multimodal alignment in human–AI interaction contexts. By jointly encoding facial expressions and gestures and introducing an explicit measure of their consistency, the UVS framework extends traditional unimodal approaches with the capacity to evaluate whether different expressive channels convey compatible affective content. This is important because inconsistencies, such as a sarcastic smile accompanied by defensive posture, are often missed by systems that consider each modality in isolation. Prior work has noted that interactions are perceived as more natural when an agent's expressions remain congruent across modalities [31], and the UVS formulation provides a principled way to operationalize this requirement.

A noteworthy advantage of the approach lies in its suitability as a feedback signal for adaptive systems. The continuous UVS score, together with its stability measure, condenses high-dimensional behavioral data into a low-dimensional indicator that can be interpreted as a proxy for coherence or engagement. Instead of relying on raw landmark trajectories or joint coordinates, an adaptive controller can monitor UVS to determine when to adjust behavior. A simple threshold can trigger changes in speaking rate, prosodic emphasis or nonverbal displays; conversely, the score can be embedded in reinforcement learning reward functions to encourage dialog policies that maintain synchrony and user comfort. Because the synchrony indices are scalar and time-resolved, they can also be examined retrospectively by researchers or practitioners to identify moments of expressive breakdown or divergence.

The notion of synchrony has implications for inclusion and personalization. For individuals who are deaf or hard of hearing, facial expressions and manual gestures form the primary channels of communication, and synchrony between these channels often signals clarity, confidence or emotional alignment. A system capable of estimating synchrony in real time could support more responsive sign-language interpretation or educational applications that adjust when a learner's expression–gesture pair becomes inconsistent. Moreover, people differ in their baseline expressiveness: some routinely display expansive gestures with smiles, whereas others use restrained body language. A personalized UVS model could learn an individual's typical face–gesture coupling and treat deviations from this learned baseline as potentially meaningful. For instance, if a normally expressive user suddenly smiles without the corresponding gestural cues, the system may infer guardedness or social constraint. Such personalized thresholds are, therefore, a promising direction for future work.

The framework also connects naturally to the concept of the affective loop. In human–human interaction, synchrony and mirroring behaviors are closely tied to rapport and empathy [19,21]. By grounding adaptation in synchrony rather than in isolated emotion labels, an agent can respond to the quality of the interaction, whether expressions are coherent, stable and mutually reinforcing—rather than merely classifying discrete affective categories. In domains such as counselling, wellbeing support or education, a synchrony-aware agent could modulate its responses based on whether a user displays stable congruence or persistent internal conflict, offering opportunities for more nuanced and contextually appropriate behavior.

Several limitations should also be acknowledged. First, $CIAS$ implicitly assumes that expressive coherence is desirable, whereas in reality intentional or context-dependent asynchrony may arise from emotional masking, politeness strategies or deliberate withholding.

In such cases, a low synchrony score is not an error but a characteristic of the interaction, and interpreting it correctly may require additional contextual signals or linguistic information. Second, the temporal scale at which synchrony is assessed is a modelling choice: while we focused on short-term dynamics, other applications may benefit from multi-scale synchrony estimation that incorporates both rapid mirroring and slower shifts in affective state. Third, our evaluation used controlled rendered data, and deploying UVS in real-world settings will face challenges related to sensor noise, occlusion, variability in camera viewpoints and heterogeneous behavioral styles.

Future work will extend the framework in several directions. One objective is to optimize the model for real-time interaction, including compression and efficient inference strategies suitable for deployment in conversational agents. Another direction is to incorporate UVS into adaptive control or reinforcement learning algorithms, allowing agents to shape their behavior so as to maintain synchrony over extended interactions. Additional research may explore synchrony-based feedback to users, e.g., in public-speaking training or educational contexts where mismatches between gesture, facial expression and verbal content can be highlighted constructively. Finally, extending the synchrony concept beyond visual cues to include voice prosody, speech rhythm or physiological signals may enable the construction of richer cross-modal synchrony indices that reflect whether tone of voice, facial expression and body movement jointly convey a coherent affective message. In this view, synchrony becomes a central organizing variable for affect-aware AI systems, enabling them to reason not only about what a user expresses, but also about how coherently this expression is realized across modalities. The real-world conversational validation demonstrates that $CIAS$ maintains high discriminative performance under structured temporal perturbations, while degrading to chance level when cross-modal structure is intentionally disrupted. This pattern confirms that the synchrony signal reflects genuine multimodal temporal coupling, rather than dataset-specific artifacts or spurious statistical correlations.

6. Conclusions

This research presented the UVS framework and the $CIAS$ as a systematic method for modeling intrapersonal visual coherence between facial emotions and body gestures. The approach considers facial and gestural streams as elements of a unified expressive system and integrates them into a common latent space, allowing for measurable and operational temporal alignment. In this form, $CIAS$ offers a continuous, time-resolved scalar that measures the extent of affective synchronization and can be incorporated into subsequent tasks, such as interaction analysis, adaptive feedback, or the regulation of responding agents.

The framework is positioned within the ESEA paradigm, which views synchrony as a fundamental resource that enhances naturalness, engagement, and interpretability in human–AI interaction. The work illustrates that synchronization can be defined as a computational attribute through the analysis of matched and mismatched multimodal sequences, rather than as a subjective heuristic. Synthetic tests demonstrate that $CIAS$ exhibits significant sensitivity to temporal coherence in optimal representational conditions. Pilot assessments suggest that synchrony measurements can act as diagnostic tools for latent representations, elucidating when visual encoders maintain temporal structure and when they diminish it. This distinction underpins the design premise of UVS: synchrony should be integrated into the representational layer rather than deduced retrospectively from ambiguous or poorly structured information.

The proposed framework provides a basis for affect-aware and synchrony-sensitive AI systems capable of reasoning not only about which signals are present, but also whether those signals jointly support a consistent affective interpretation. Future research will

extend UVS to richer multimodal settings by incorporating prosody, lexical cues or physiological signals; by developing learned cross-modal encoders trained through contrastive synchrony objectives; and by adopting structured 3D representations (e.g., SMPL-X and FLAME) for stable modelling of face–body dynamics. Long-term interaction scenarios further motivate personalized synchrony baselines and multi-scale temporal analysis. Overall, unified visual synchrony offers a principled foundation for creating AI systems that track the coherence of expressive channels, aligning computational processing more closely with the holistic mechanisms through which humans interpret one another's behavior. Collectively, these findings position *CIAS* as an innovative and comprehensible metric of visual synchrony, apart from current similarity-based or task-specific multimodal representations.

Author Contributions: Conceptualization, Y.S. and S.K.; methodology, Y.S., software, S.K. and B.S.; validation, A.Y. and E.D.; formal analysis, Y.S.; investigation, D.T.; resources, E.D.; data curation, B.S.; writing—original draft preparation, A.Y.; writing—review and editing, S.K. and A.Y.; visualization, E.D. and D.T.; supervision, S.K. and M.T.; project administration, E.D. and M.T.; funding acquisition, 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 original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding authors.

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

References

1. Baltrušaitis, T.; Ahuja, C.; Morency, L.-P. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 423–443. [CrossRef] [PubMed]
2. Atrey, P.K.; Hossain, M.A.; El Saddik, A.; Kankanhalli, M.S. Multimodal fusion for multimedia analysis: A survey. *Multimed. Syst.* **2010**, *16*, 345–379. [CrossRef]
3. Zadeh, A.; Liang, P.P.; Mazumder, N.; Poria, S.; Morency, L.-P. Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion. In *Proceedings of ACL*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; pp. 2236–2246. [CrossRef]
4. Alexanderson, S.; Henter, G.E.; Kucherenko, T.; Beskow, J. Style-Controllable Speech-Driven Gesture Synthesis Using Normalising Flows. *Comput. Graph. Forum* **2020**, *39*, 487–496. [CrossRef]
5. Liu, H.; Zhu, Z.; Becherini, G.; Peng, Y.; Su, M.; Zhou, Y.; Zhe, X.; Iwamoto, N.; Zheng, B.; Black, M.J. EMAGE: Towards Unified Holistic Co-Speech Gesture Generation via Expressive Masked Audio Gesture Modeling. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2024; pp. 1144–1154. [CrossRef]
6. Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Narayanan, S. IEMOCAP: Interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* **2008**, *42*, 335–359. [CrossRef]
7. Tsai, Y.-H.H.; Bai, S.; Yamada, M.; Morency, L.-P.; Salakhutdinov, R. Multimodal Transformer for Unaligned Multimodal Language Sequences. In *Proceedings of ACL*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 6558–6569. [CrossRef]
8. Praveen, R.G.; Alam, J. Recursive Joint Cross-Modal Attention for Multimodal Fusion in Dimensional Emotion Recognition. *arXiv* **2024**, arXiv:2403.13659. [CrossRef]
9. Waligora, P.; Aslam, H.; Zeeshan, O.; Koerich, A.; Pedersoli, M.; Granger, E. Joint Multimodal Transformer for Emotion Recognition in the Wild. *arXiv* **2024**, arXiv:2403.10488. [CrossRef]
10. Plutchik, R. A Psychoevolutionary Theory of Emotions. *Soc. Sci. Inf.* **1982**, *21*, 529–553. [CrossRef]
11. Yernar, S. Multimodal Emotion and Gesture: UVS/*CIAS* Implementation and Evaluation Scripts. GitHub Repository. 2026. Available online: https://github.com/ernar65019920/Multimodal_emoution_and_gesture (accessed on 15 November 2025).
12. McNeill, D. *Hand and Mind: What Gestures Reveal About Thought*; University of Chicago Press: Chicago, IL, USA, 1992.

13. Wang, Y.; Neff, M. The Influence of Prosody on the Requirements for Gesture–Text Alignment. In *Intelligent Virtual Agents (IVA 2013)*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 180–188. [CrossRef]
14. Kendon, A. *Gesture: Visible Action as Utterance*; Cambridge University Press: Cambridge, UK, 2004.
15. Grewe, C.M.; Fehrenbach, J.; Hornecker, E. Statistical Learning of Facial Expressions Improves the Perceived Realism of Virtual Characters. *Front. Virtual Real.* **2021**, *2*, 619811. [CrossRef]
16. Condon, W.S.; Ogston, W.D. A segmentation of behavior. *J. Psychiatr. Res.* **1967**, *5*, 221–235. [CrossRef]
17. Fu, D.; Liu, Y.; Delaherche, E.; Chetouani, M. Interpersonal Physiological Synchrony for Detecting Social Interaction Quality. *Front. Psychol.* **2021**, *12*, 749710. [CrossRef]
18. Wu, Y.; Zhang, L.; Chen, H.; Li, Y. A Comprehensive Review of Multimodal Emotion Recognition. *Biomimetics* **2025**, *10*, 418. [CrossRef]
19. Mehrabian, A. *Silent Messages*; Wadsworth: Belmont, CA, USA, 1971.
20. de Gelder, B. Towards the neurobiology of emotional body language. *Nat. Rev. Neurosci.* **2006**, *7*, 242–249. [CrossRef]
21. Cowen, A.S.; Keltner, D. What the Face Displays: Mapping 28 Emotions Conveyed by Naturalistic Expression. *Am. Psychol.* **2020**, *75*, 349–364. . [CrossRef]
22. 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]
23. Aviezer, H.; Hassin, R.R.; Ryan, J.; Grady, C.; Susskind, J.; Anderson, A.; Moscovitch, M.; Bentin, S. Angry, disgusted, or afraid? Studies on the malleability of emotion perception. *Psychol. Sci.* **2008**, *19*, 724–732. [CrossRef] [PubMed]
24. Kret, M.E.; de Gelder, B. Social context influences recognition of bodily expressions. *Exp. Brain Res.* **2010**, *203*, 169–180. [CrossRef] [PubMed]
25. Pavlova, M.A. Biological motion processing as a hallmark of social cognition. *Cereb. Cortex* **2012**, *22*, 981–995. [CrossRef] [PubMed]
26. Argyle, M. *Bodily Communication*, 2nd ed.; Methuen: London, UK, 1988.
27. Harrigan, J.A.; Rosenthal, R.; Scherer, K.R. *The New Handbook of Methods in Nonverbal Behavior Research*; Oxford University Press: Oxford, UK, 2005.
28. Scherer, K.R. The dynamic architecture of emotion: Evidence for the component process model. *Cogn. Emot.* **2009**, *23*, 1307–1351. [CrossRef]
29. Liu, Y.; Wang, C.; Lu, M.; Yang, J.; Gui, J.; Zhang, S. From Simple to Complex Scenes: Learning Robust Feature Representations for Accurate Human Parsing. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 5449–5462. [CrossRef] [PubMed]
30. Wang, C.; Zhang, Q.; Wang, X.; Zhou, L.; Li, Q.; Xia, Z.; Ma, B.; Shi, Y.-Q. Light-Field Image Multiple Reversible Robust Watermarking Against Geometric Attacks. *IEEE Trans. Dependable Secur. Comput.* **2025**, *22*, 5861–5875. [CrossRef]
31. Ekman, P.; Friesen, W. *Facial Action Coding System (FACS)*; Consulting Psychologists Press: Palo Alto, CA, USA, 1978.

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.