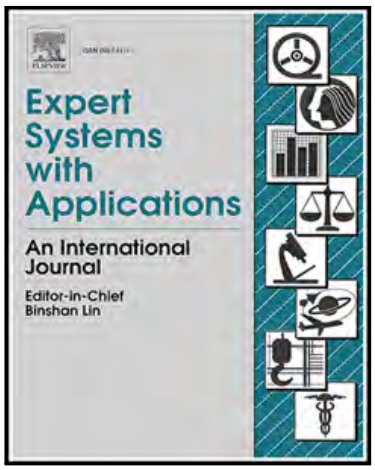
Journal Pre-proof



Decoupled Semantic Alignment and Emotion-Aware Gated Fusion for Group Emotion Recognition

Hongxi Yi, Yufei Xie, Dong Zhang, Wenxing Liu, Ming Li, Dah-Jye Lee

PII: S0957-4174(26)03550-5
DOI: https://doi.org/10.1016/j.eswa.2026.134646
Reference: ESWA 134646

To appear in: *Expert Systems With Applications*

Received date: 8 May 2026
Revised date: 18 August 2026
Accepted date: 4 October 2026

Please cite this article as: Hongxi Yi, Yufei Xie, Dong Zhang, Wenxing Liu, Ming Li, Dah-Jye Lee, Decoupled Semantic Alignment and Emotion-Aware Gated Fusion for Group Emotion Recognition, *Expert Systems With Applications* (2025), doi: https://doi.org/10.1016/j.eswa.2026.134646

This is a PDF of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability. This version will undergo additional copyediting, typesetting and review before it is published in its final form. As such, this version is no longer the Accepted Manuscript, but it is not yet the definitive Version of Record; we are providing this early version to give early visibility of the article. Please note that Elsevier's sharing policy for the Published Journal Article applies to this version, see: https://www.elsevier.com/about/policies-and-standards/sharing#4-published-journal-article. Please also note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2026 Published by Elsevier Ltd.

Journal Pre-proof

Highlights

Decoupled Semantic Alignment and Emotion-Aware Gated Fusion for Group Emotion Recognition

Hongxi Yi, Yufei Xie, Dong Zhang, Wenxing Liu, Ming Li, Dah-Jye Lee

- A modality-decoupled language-guided framework is proposed for video-based group emotion recognition.
- Psychology-informed visual and acoustic narratives reduce semantic entanglement across modalities.
- SETA and SEGF improve temporal localization and channel-wise multimodal fusion.
- Across 10 runs, the method achieves 77.69±1.04% on VGAF and 97.87±0.47% on GECV, with best-run accuracies of 80.03% and 98.52%.

Journal Pre-proof

Journal Pre-proof

Decoupled Semantic Alignment and Emotion-Aware Gated Fusion for Group Emotion Recognition

Hongxi Yi[a], Yufei Xie[a], Dong Zhang[a,*], Wenxing Liu[b], Ming Li[c,b] and Dah-Jye Lee[d]

[a]School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou, 510006, Guangdong, China
[b]School of Computer Science, Wuhan University, Wuhan, 430072, Hubei, China
[c]School of Artificial Intelligence, The Chinese University of Hong Kong, Shenzhen, Shenzhen, 518172, Guangdong, China
[d]Department of Electrical and Computer Engineering, Brigham Young University, Provo, 84602, Utah, USA

ARTICLE INFO

Keywords:
Group emotion recognition
Vision-language models
Contrastive alignment
Gated fusion
Affective computing

ABSTRACT

Video-based group emotion recognition must resolve three coupled problems: semantic entanglement among multimodal cues, the loss of transient affective climaxes during temporal aggregation, and the propagation of noisy sensory features during fusion. We propose a Modality-Decoupled Language-Guided Framework that uses a vision-language model as an offline semantic engine to generate independent psychology-informed visual and acoustic narratives. The resulting modality-specific text anchors are aligned with CLIP visual features and CLAP acoustic features through a dual InfoNCE objective. A Soft Emotion-Aware Temporal Attention (SETA) module predicts a continuous emotion distribution and maps it to an affective query for frame-wise temporal weighting. The same distribution is projected into a fusion-specific query for Soft Emotion-Aware Gated Fusion (SEGF), which performs complementary channel-wise routing of the refined visual and acoustic features. The main novelty is the coordinated use of narrative-level modality decoupling and a shared soft affective hypothesis across both temporal localization and multimodal fusion. Across 10 independent runs, the framework achieves 77.69±1.04% mean accuracy on VGAF and 97.87±0.47% on GECV, with best-run accuracies of 80.03% and 98.52%, respectively. Interpretability, corruption, efficiency, narrative-decoupling, and failure-case analyses further characterize the behavior of the proposed modules.

1. Introduction

Understanding the overall mood of a group of people requires moving beyond the analysis of individual facial expressions to interpreting the complex interactions among multiple participants. The main objective of Group Emotion Recognition (GER) is to infer the shared emotional state of a group from diverse sources of data. Compared with individual emotion recognition, GER must reason about collective phenomena such as emotional contagion, shared attention, and socially coordinated behaviors that propagate among multiple individuals [1]. Reliable GER serves as a core engine for intelligent urban monitoring, sociological analysis, and human-centered multimedia analytics, where decisions depend heavily on the perceived overall atmosphere rather than on any single person's facial expression.

Moving from static images to continuous video sequences provides important temporal information for group emotion analysis. Nevertheless, transforming these raw and heterogeneous signals into reliable emotion predictions is hindered by three fundamental structural weaknesses. First, traditional methods for learning emotion representation focus on raw pixel- and waveform-level patterns and therefore tend to overlook broader social context. Without such crucial semantic context, a network may misinterpret a hurried evacuation as a joyful celebration, since both may exhibit high-energy and chaotic low-level features. Second, understanding the temporal evolution of group behavior requires the accurate identification of emotional turning points. Traditional sequence summarization methods usually compress these dynamic changes into a single flat vector, thereby losing important short-term peaks in crowd reactions. Third, existing methods still rely on rigid multimodal fusion schemes. Under real-world environmental corruptions, such as severe camera occlusion or strong background interference, these models cannot adaptively assess signal quality. As a result, they may indiscriminately fuse low-quality signals, which distorts the final representation and reduces overall accuracy.

To introduce explicit cognitive reasoning into deep networks, recent studies have begun to leverage vision-language models (VLMs) to generate textual descriptions for GER videos, thereby incorporating rich prior knowledge into the reasoning process. However, this strategy introduces a critical new limitation: semantic entanglement in the language domain. Existing methods typically prompt the VLM to generate a single holistic caption summarizing the entire video. Our preliminary exploration suggests that such a unified description entangles visual and acoustic cues, making it inherently difficult to align with modality-specific encoders such as CLIP and CLAP. If the quality of one modality is degraded, for example, when the audio is unclear, the unified caption may implicitly rely on information from the higher-quality modality, leading to cross-modal hallucination. Moreover, compressing complex video dynamics into a single fixed sentence severely limits the model's ability to

*Corresponding author. Address: School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou, 510006, China.
yihx6@mail2.sysu.edu.cn (H. Yi); xieyf63@mail2.sysu.edu.cn (Y. Xie); zhangd@mail.sysu.edu.cn (D. Zhang); wenxingliu@whu.edu.cn (W. Liu); mingli369@cuhk.edu.cn (M. Li); djlee@byu.edu (D. Lee)
ORCID(s):

Journal Pre-proof

support fine-grained temporal selection and robust feature fusion. Existing studies therefore address language guidance, temporal modeling, and multimodal fusion largely as separate problems; they do not jointly prevent narrative-level cross-modal leakage, emotion-agnostic temporal smoothing, and channel-wise noise propagation.

This failure mode is observable rather than merely hypothetical. In the diagnostic study presented in Section 4.8, we merge the visual and acoustic prompt instructions and use GLM-4.6V-Flash to generate one unified audiovisual description, which is then supplied as both the visual-text and audio-text input. Across all 766 VGAF validation samples, this shared-description condition reduces accuracy from 80.03% to 76.50%. For sample 209_15, the shared description flips a correct *Negative* prediction into *Positive* while the raw sensory inputs remain unchanged. This example motivates preserving modality-specific semantic anchors until the final fusion stage.

To address these tightly coupled challenges, we propose a Modality-Decoupled Language-Guided Framework for video-based group emotion recognition. To overcome semantic entanglement, we introduce a narrative-level decoupling strategy. Instead of generating a single mixed caption, we use two psychology-informed prompts to guide a pre-trained VLM to produce two independent descriptions: a visual narrative focusing on behaviors and context, and an acoustic narrative focusing on tone and rhythm. These decoupled texts serve as clean semantic anchors and are aligned with visual and audio features through a dual InfoNCE loss. To address the temporal modeling and fusion limitations, we further introduce the concept of "soft emotion." Unlike traditional fixed emotion labels, which rely on rigid one-hot classification and ignore the uncertainty inherent in crowd emotions, a soft emotion prototype represents a continuous probability distribution over multiple emotional states. Based on this idea, we design a Soft Emotion-Aware Temporal Attention (SETA) module, which uses visual-text semantics to predict a soft emotion distribution that subsequently acts as a dynamic query to localize emotional peaks across video frames. Finally, we propose a Soft Emotion-Aware Gated Fusion (SEGF) strategy that is conditioned on this global soft-emotion prior to dynamically reweight and fuse reliable feature channels from both the visual and audio streams. The originality of the proposed solution lies in using the same continuous affective hypothesis to connect temporal localization and multimodal routing, while keeping visual and acoustic language priors explicitly separated until the final fusion stage.

The main contributions of this work are summarized as follows:

- We identify the key problem of semantic entanglement in VLM-driven GER and propose a narrative-level modality-decoupling scheme. By generating clearly separated visual and acoustic text anchors and combining them with dual InfoNCE alignment, we provide clean and fine-grained cross-modal guidance while mitigating cross-modal hallucination.
- We design a Soft Emotion-Aware Temporal Attention (SETA) module for interpretable temporal modeling. By using a predicted soft emotion distribution rather than fixed hard labels as an active query, SETA alleviates temporal flattening and accurately localizes emotional peaks in an interpretable, hypothesis-driven manner.
- We propose a Soft Emotion-Aware Gated Fusion (SEGF) mechanism that uses the soft-emotion hypothesis to suppress channel-wise noise and integrate reliable modality information. Extensive experiments on the VGAF and GECV benchmarks demonstrate the effectiveness and strong performance of the proposed framework.

2. Related Work

This section reviews three lines of research that are most relevant to the proposed framework: group-level representation learning, temporal attention and multimodal fusion, and language-based knowledge injection. The review is organized to identify the specific gaps addressed by narrative decoupling, SETA, and SEGF.

2.1. Representation Learning for Group Emotion

Estimating crowd emotion goes beyond simply aggregating individual facial expressions; it requires decoding the emergent atmosphere shaped by social interaction and affective contagion [1, 2]. Historically, the extraction of group-level features relied heavily on static photographs to capture spatial arrangements. Early architectures employed multi-stream or hierarchical convolutional neural networks (CNNs) to concurrently process localized facial regions and broader environmental contexts [3–5]. Nagarajan and Oruganti further investigated image-based GER under adverse face-detection conditions, showing that global scene and place features can remain competitive when faces are occluded, non-frontal, or too small to detect reliably [6]. Subsequent advances introduced graph neural networks (GNNs) to model the social structure of a group by representing participants as interconnected nodes and learning their relational dependencies [7, 8]. Such graph-based context modeling paradigms have also proven effective in capturing multi-agent interactions in broader affective computing tasks, such as emotion recognition in conversations [9]. To address the intrinsic ambiguity and subjective annotation noise in naturalistic crowd images, various uncertainty-modeling and label-distribution learning strategies were subsequently developed [10, 11].

Nevertheless, relying solely on still frames overlooks the crucial temporal flow and acoustic context of collective events. Researchers transitioned to video inputs to capture this dynamic progression, utilizing 3D spatiotemporal convolutions and dynamic graph structures to trace behavioral changes over time [12, 13]. Concurrently, the incorporation of audio streams, which encode cheers, speech, and background noise through deep spectrogram embeddings,

Journal Pre-proof

provided a valuable complementary signal, particularly under severe visual occlusion [14, 15]. Despite the transition toward spatiotemporal and multisensory architectures, these feature extractors largely remain opaque and purely data-driven. By attempting to map raw inputs directly to emotion labels without exploiting structured high-level semantic guidance, they still struggle to capture the complex psychological underpinnings of crowd behavior [16].

2.2. Attention and Fusion Mechanisms in Multimodal GER

Because video sequences contain highly heterogeneous sensory streams, the fusion strategy used to integrate these modalities largely determines overall system performance. Early audiovisual approaches often relied on simple late concatenation or linear score fusion, implicitly assuming that audio and visual channels provide equally reliable evidence across different contexts [17, 18]. To handle the varying reliability of heterogeneous streams, recent studies have explored modality-adaptive architectures that dynamically regulate sensory contributions [19]. Building on the need for deeper synergistic interactions, more recent frameworks have incorporated cross-attention mechanisms [20, 21]. Such designs enable mutual conditioning between visual and acoustic feature spaces, producing context-aware embeddings similar to those used in complex multimodal sequence modeling [22].

Despite these advances, modern integration techniques still suffer from two major structural limitations. First, temporal pooling operations remain largely agnostic to the target emotional state. By compressing entire video clips into a single global descriptor through standard pooling or conventional self-attention, existing methods fail to use an affective query to actively identify emotionally salient moments [23]. Second, conventional fusion algorithms typically assign only a single scalar weight to each modality, overlooking the need for fine-grained dimension-level modulation [24]. In practice, different feature dimensions encode distinct semantics, such as visual facial contours and auditory pitch variations. Without a channel-specific gating mechanism driven by global affective context, current networks are vulnerable to transient noise, temporal misalignment, and conflicting multimodal cues.

2.3. Knowledge Injection and Cross-Modal Alignment

Large-scale pretraining of vision–language models has substantially reshaped cross-modal representation learning. Dual-encoder architectures such as CLIP and CLAP construct shared embedding spaces in which visual and acoustic modalities are aligned with linguistic concepts, thereby facilitating robust zero-shot applications and prompt-based inference [25, 26]. Going one step further, generative multimodal models, including GPT-4V and GLM-4.6V-Flash, have demonstrated strong zero-shot reasoning ability in affective-context understanding [27, 28]. As a result, researchers have increasingly adopted contrastive objectives such as InfoNCE, together with more advanced prototype-guided contrastive training strategies [29], to bridge the gap between low-level visual or acoustic signals and high-level linguistic priors, thereby learning more robust emotion representations [30, 31].

Deploying these large-scale networks as direct inference engines for crowd affect remains computationally prohibitive and susceptible to domain discrepancies. Instead, a practical approach involves utilizing these massive models strictly as offline annotators to create rich semantic targets for training smaller, task-specific models. However, current methodologies in this area mainly rely on extracting a single textual summary for the entire audiovisual clip. Our experiments show that this unified approach inherently leads to narrative blending. By constraining the model to a single output, the generative model merges sound and visual information; consequently, when the audio is ambiguous, the text may incorrectly infer sound events solely from visual cues. To solve this problem, we introduce a fully decoupled prompting mechanism. We build a clean and accurate semantic basis for crowd emotion analysis by generating separate textual descriptions for the visual scene and the auditory environment, and then linking them to their corresponding features through a dual-contrastive framework.

Table 1 summarizes how the proposed framework is positioned relative to representative recent GER studies. The comparison is intended to clarify that our contribution is not a replacement for pose, graph, or transformer representations; rather, it adds modality-specific semantic anchoring and emotion-conditioned routing that are absent from these paradigms.

3. Methodology

This section first presents the overall pipeline and then formalizes each unit in the order in which information flows through the network: decoupled narrative generation, modality-specific encoding, soft emotion-aware temporal modeling, gated multimodal fusion, and joint optimization. Unlike conventional architectures that rely solely on bottom-up sensory processing, our framework adopts a top-down, concept-guided paradigm to reduce the inherent ambiguity of crowd emotion analysis. As illustrated in Fig. 1, the proposed architecture consists of three collaborative stages.

First, we establish modality-decoupled semantic anchors by generating separate visual and acoustic narratives with a large vision-language model (VLM). Second, these narratives serve as prior knowledge to guide modality-specific encoding and contrastive alignment via a dual InfoNCE loss, ensuring that both visual and acoustic features are projected into an emotion-discriminative space. Third, instead of employing conventional late fusion, we introduce an affective hypothesis-driven aggregation strategy. This strategy includes a Soft Emotion-Aware Temporal Attention module to aggregate frame-level cues and a Soft Emotion-Aware Gated Fusion mechanism to dynamically modulate modality

Journal Pre-proof

Table 1
Comparative positioning of representative recent group emotion recognition studies. The limitations are stated specifically with respect to the three problems targeted in this work: semantic entanglement, emotion-agnostic temporal aggregation, and inflexible multimodal fusion.

Study or paradigm	Main design	Limitation relevant to this study	How the proposed framework addresses it
Slogrove and Van Der Haar [32]	Pose estimation with LSTM temporal modeling	Focuses on body pose and a selected evaluation subset; appearance, audio, and language-level social context are not jointly modeled.	Uses the official VGAF validation protocol and combines visual, acoustic, and modality-specific textual evidence.
Quach et al. [17]	Non-volume-preserving audiovisual fusion	Learns a global fusion mapping but does not explicitly separate visual and acoustic narratives or use an affective query for frame selection.	Introduces decoupled semantic anchors, SETA temporal localization, and soft-emotion-conditioned channel gates.
Augusma et al. [33]	Privacy-compliant multimodal features	Achieves a strong VGAF result, but does not provide language-guided modality decoupling or interpretable emotion-conditioned routing.	Adds explicit semantic alignment and exposes both temporal attention weights and modality gate statistics.
Wang et al. [13]	Spatial-temporal graph convolution	Models relational dynamics effectively, but remains primarily bottom-up and lacks modality-specific linguistic priors and channel-wise noise suppression.	Uses VLM-derived semantic anchors and SEGF to modulate visual and acoustic channels according to a global affective hypothesis.
Huang and Xu [16]	Multimodal spatiotemporal semisupervised transformer	Improves data utilization but does not directly address unified-caption leakage or soft-prototype-guided temporal and fusion decisions.	Separates narrative generation by modality and couples the same soft prototype to both SETA and SEGF.
Proposed framework	Decoupled VLM narratives, dual InfoNCE, SETA, and SEGF	Depends on pretrained encoders and offline VLM descriptions; external cross-dataset generalization remains to be established.	Provides explicit, interpretable semantic alignment and adaptive temporal/channel routing, with limitations analyzed in Section 4.11.

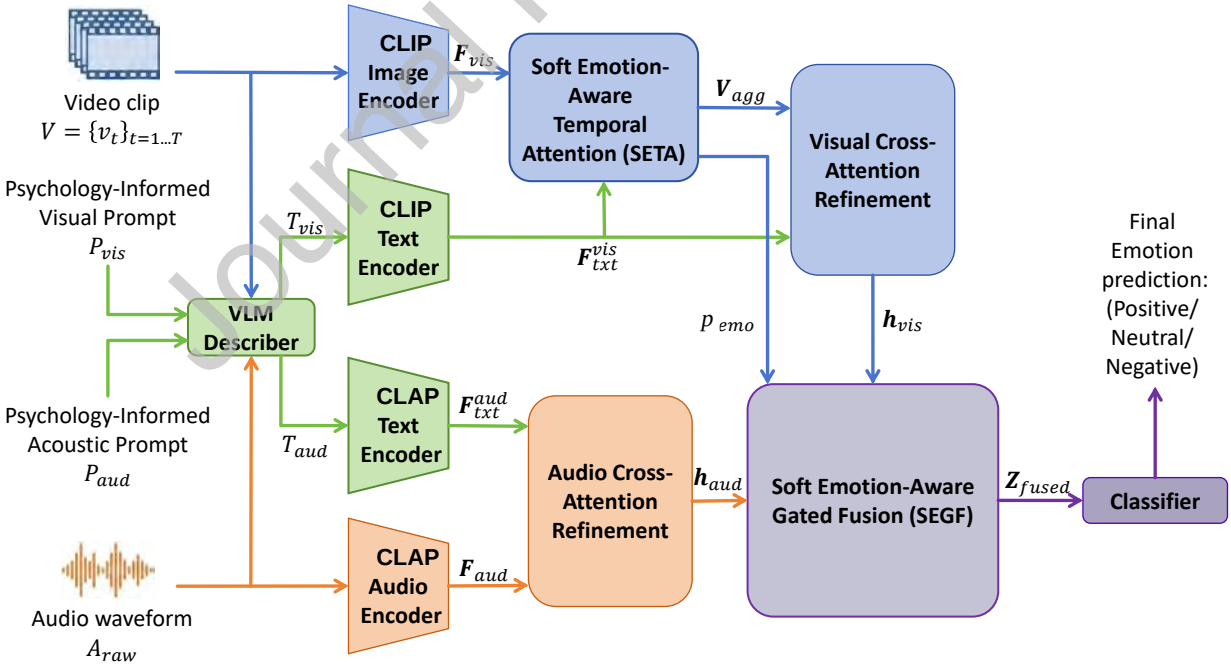


Figure 1: Overall architecture of the proposed Modality-Decoupled Language-Guided Framework. The visual (blue) and acoustic (orange) sensory streams are paired with independently generated psychology-informed narratives (green). SETA produces a context-aware video embedding and a soft affective probability, while cross-attention refinement and SEGF use these semantic cues to obtain a channel-wise fused representation for three-class group emotion prediction.

Journal Pre-proof

Table 2
Architectural differences between representative temporal/fusion mechanisms and the proposed SETA and SEGF units.

Mechanism	Conditioning signal	Temporal operation	Fusion granularity	Distinctive capability or limitation
Temporal average pooling	None	Uniform averaging over frames	Not applicable	Efficient but can suppress short affective climaxes.
Generic self/cross-attention [22]	Feature content and learned projections	Content-dependent token interaction	Token or representation level	Captures dependencies but is not explicitly conditioned on a predicted group-emotion hypothesis.
Non-volume-preserving fusion [17]	Joint audiovisual representation	Depends on upstream temporal encoder	Global learned fusion	Models multimodal dependence, but does not use modality-decoupled language anchors or a shared soft affective query.
SETA (ours)	Soft emotion prototype predicted from baseline visual context and visual text	Emotion-conditioned frame-wise softmax weighting	Not applicable	Localizes salient frames while retaining uncertainty through a continuous distribution.
SEGF (ours)	Refined visual/audio features and a fusion-specific soft affective query	Uses the SETA-aggregated visual representation	Per feature channel	Produces complementary visual/audio gates and can suppress unreliable dimensions rather than assigning one scalar per modality.

contributions based on a predicted global emotion distribution. By shifting the paradigm from label-driven pooling to language-guided reasoning, our framework achieves improved effectiveness and robustness in collective emotion analysis.

Table 2 explicitly distinguishes the proposed mechanisms from commonly used temporal and fusion operators. SETA is different from generic attention because its query is a predicted continuous emotion prototype rather than only the current feature content; SEGF is different from scalar modality weighting because it makes a complementary decision for every feature channel.

3.1. Decoupled Language Knowledge Generation

The core idea of our approach is to reformulate emotion classification as a conceptually anchored mapping problem. Although standard convolutional and transformer backbones are effective at identifying low-level audiovisual patterns, they are inherently limited in capturing broader sociological context. Vision-language models (VLMs), such as GLM-4.6V-Flash, possess broad world knowledge and strong inferential capabilities. To alleviate the limitations caused by cross-domain discrepancy and sparse annotations in GER datasets, we employ the VLM as an offline linguistic annotator to generate independent, theoretically grounded narratives.

To avoid the narrative entanglement frequently caused by mixed-modality prompting, we abandon generic querying strategies. Instead, we design a strictly decoupled protocol in which the extraction criteria for both visual and audio streams are explicitly guided by canonical theories in social psychology.

(1) Psychology-informed visual narrative. The visual manifestation of group mood is inherently multidimensional. Grounded in the universality of facial expressions [34] and emotional contagion theory [1], the visual basis of crowd emotion lies in expressive behaviors that propagate among participants. Extending this perspective from the individual to the collective level, intergroup conflict theory [35] and the ripple effect [2] explain how large-scale social structures and coordinated actions shape overall group sentiment. Moreover, cultural universality [36] and context-dependent emotion perception [37] suggest that environmental cues and social norms strongly influence how such emotional states are interpreted. Guided by these principles, the visual prompt P_{vis} instructs the VLM to analyze the visual stream from four aspects:

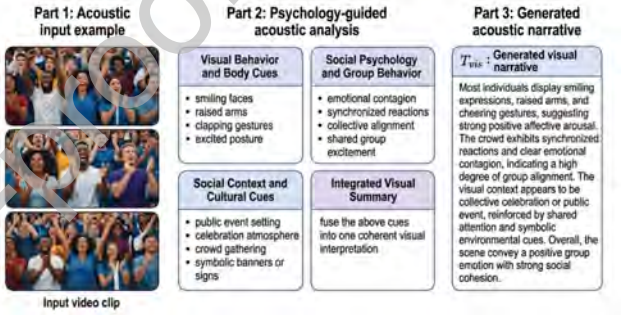


Figure 2: Example of psychology-informed visual narrative generation. The prompt separates observable behavior, group interaction, and social context so that the resulting text serves as a visual-only semantic anchor.

(i) **Visual behavior and body cues:** capturing facial movements and bodily gestures at the individual level;

(ii) **Social psychology and group behavior:** identifying macroscopic interpersonal dynamics, including collective alignment or opposition;

(iii) **Social context and cultural cues:** inferring the nature of the gathering and relevant social symbols;

(iv) **Integrated visual summary:** synthesizing the above observations into an independent behavioral narrative.

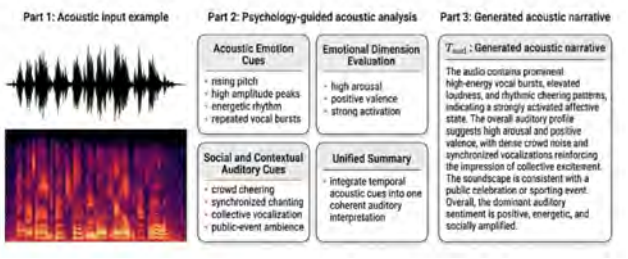


Figure 3: Example of psychology-informed acoustic narrative generation. The prompt focuses on prosody, valence–arousal cues, crowd sounds, and auditory context without borrowing visual evidence.

(2) Psychology-informed acoustic narrative. Similarly, the acoustic expression of group emotion follows interpretable psychological regularities. Research on acoustic profiles in vocal expression [38] links fundamental acoustic attributes, such as vocal frequency, amplitude, and rhythm, to specific emotional categories. For deeper cognitive interpretation, the circumplex model of affect [39] and the component process model [40] provide a theoretical basis for analyzing arousal, valence, and cognitive appraisal. In addition, interaction ritual chain theory [41] highlights the crucial role of synchronized chanting and collective vocalization in shaping crowd energy. Guided by these principles, the acoustic prompt P_{aud} instructs the VLM to analyze the audio stream from four complementary perspectives:

(i) **Acoustic emotion cues:** profiling pitch fluctuations, volume peaks, and rhythmic pacing;

(ii) **Emotional dimension evaluation:** characterizing the auditory signal along the valence and arousal dimensions;

(iii) **Social and contextual auditory cues:** describing the background ambiance and synchronized crowd noise;

(iv) **Unified summary:** generating a chronologically coherent summary of the dominant auditory sentiment.

Mathematically, given a video clip V and its corresponding raw audio waveform A_{raw}, the narrative generation process is defined as

$$T_{\text{vis}} = \text{VLM}(V, P_{\text{vis}}), \qquad T_{\text{aud}} = \text{VLM}(A_{\text{raw}}, P_{\text{aud}}), \quad (1)$$

where T_{vis} and T_{aud} denote the generated visual and acoustic narratives, respectively. These independent texts are subsequently embedded by the CLIP and CLAP text encoders to obtain the semantic feature vectors $\mathbf{F}_{\text{txt}}^{\text{vis}} \in \mathbb{R}^D$ and $\mathbf{F}_{\text{txt}}^{\text{aud}} \in \mathbb{R}^D$. These embeddings serve as robust semantic priors for the downstream attention and aggregation modules.

3.2. Modality-Specific Encoding

To avoid the high computational cost of training feature extractors from scratch, we leverage the embedding spaces learned by large-scale contrastive pretraining models, which naturally align sensory inputs with linguistic semantics.

Visual and semantic encoding via CLIP. We adopt the Contrastive Language–Image Pre-training (CLIP) model [25] as the backbone for visual-semantic feature extraction. Given a sequence of T video frames, the CLIP image encoder maps each frame v_t into a temporal feature sequence

$$\mathbf{F}_{\text{vis}} = \left\{\mathbf{f}_{\text{vis}}^1, \mathbf{f}_{\text{vis}}^2, \ldots, \mathbf{f}_{\text{vis}}^T\right\}, \qquad \mathbf{F}_{\text{vis}} \in \mathbb{R}^{T\times D}, \quad (2)$$

where

$$\mathbf{f}_{\text{vis}}^t = \text{Enc}_{\text{img}}^{\text{clip}}(v_t). \quad (3)$$

Meanwhile, the visual narrative T_{vis} generated in Section 3.1 is fed into the CLIP text encoder, yielding a global visual-semantic representation

$$\mathbf{F}_{\text{txt}}^{\text{vis}} = \text{Enc}_{\text{txt}}^{\text{clip}}(T_{\text{vis}}), \quad (4)$$

which captures the high-level social and psychological context of the visual scene.

Auditory and semantic encoding via CLAP. For acoustic modality, we employ the Contrastive Language–Audio Pretraining (CLAP) model [26] for feature extraction. The raw audio waveform A_{raw} is first resampled to 48 kHz and then encoded by the CLAP audio encoder into a global acoustic representation $\mathbf{F}_{\text{aud}} \in \mathbb{R}^D$:

$$\mathbf{F}_{\text{aud}} = \text{Enc}_{\text{aud}}^{\text{clap}}(A_{\text{raw}}). \quad (5)$$

Similarly, the acoustic narrative T_{aud} is embedded by the CLAP text encoder to obtain the acoustic-semantic representation

$$\mathbf{F}_{\text{txt}}^{\text{aud}} = \text{Enc}_{\text{txt}}^{\text{clap}}(T_{\text{aud}}). \quad (6)$$

Because CLIP and CLAP follow similar contrastive pretraining principles, we denote the four resulting representations by

$$\mathcal{F} = \{\mathbf{F}_{\text{vis}}, \mathbf{F}_{\text{txt}}^{\text{vis}}, \mathbf{F}_{\text{aud}}, \mathbf{F}_{\text{txt}}^{\text{aud}}\}.$$

All elements of $\mathcal{F}$ lie in a compatible D-dimensional space, where $D = 512$. This dimensional consistency allows the generated linguistic anchors to directly support downstream temporal localization and cross-modal information routing without requiring complex intermediate adaptation modules.

3.3. Soft Emotion-Aware Temporal Modeling

Unlike generic temporal pooling techniques, such as recurrent networks and standard self-attention, which process sequences without considering task-specific emotional hypotheses, our framework recognizes that emotional evidence

Journal Pre-proof

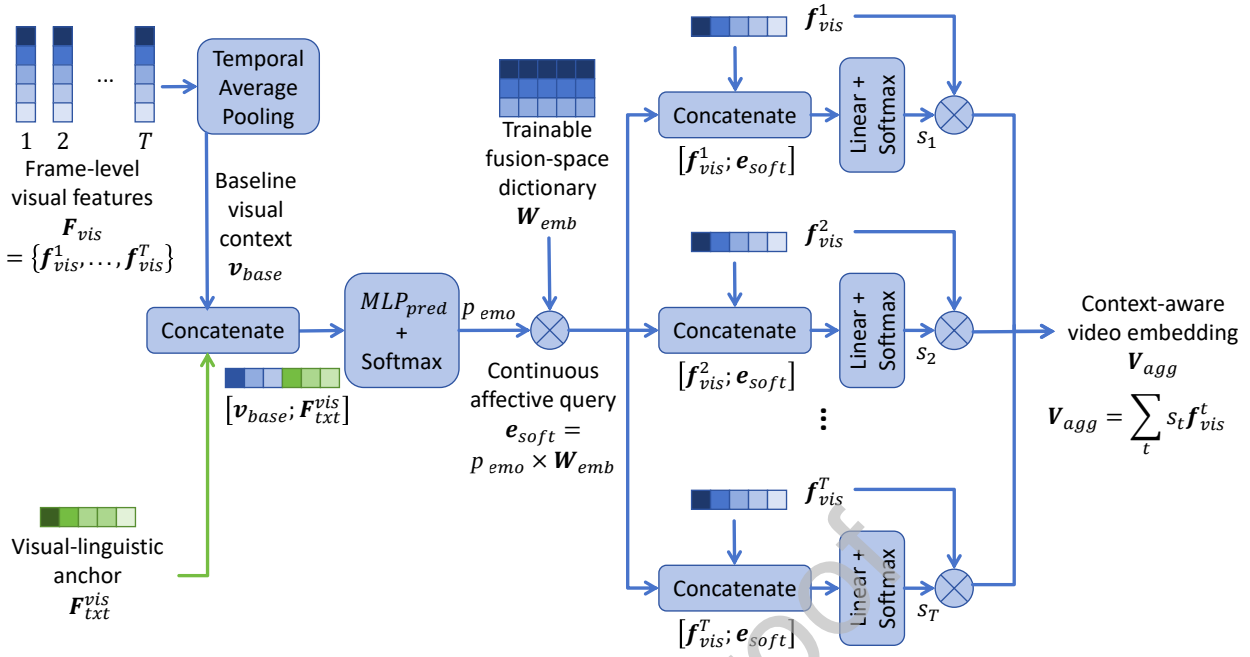


Figure 4: Architecture of the Soft Emotion-Aware Temporal Attention (SETA) module. A baseline visual context and the visual-linguistic anchor predict a soft emotional probability distribution. The distribution is mapped through a trainable dictionary to a continuous affective query, concatenated with every frame descriptor, and converted into normalized temporal weights whose weighted sum yields the context-aware video embedding $\mathbf{V}_{\text{agg}}$.

is unevenly distributed over time. In dynamic crowd scenes, transient outbursts, such as an abrupt collective celebration, often encode the most salient affective information. To actively identify these critical segments, we design the Soft Emotion-Aware Temporal Attention (SETA) module, which leverages text-guided affective priors to direct temporal aggregation.

First, an unweighted temporal average is computed to form a baseline visual context $\mathbf{v}_{\text{base}} \in \mathbb{R}^D$ from the frame-level feature sequence $\mathbf{F}_{\text{vis}} = \{\mathbf{f}_{\text{vis}}^1, \ldots, \mathbf{f}_{\text{vis}}^T\}$:

$$\mathbf{v}_{\text{base}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{f}_{\text{vis}}^t. \tag{7}$$

To infuse this baseline with semantic awareness, it is concatenated with the visual-semantic anchor $\mathbf{F}_{\text{txt}}^{\text{vis}}$ obtained in Section 3.2. The resulting representation is then fed into a prediction multilayer perceptron (MLP), composed of layer normalization, a 256-dimensional hidden projection, a GELU activation, dropout with a rate of 0.1, and a linear classifier, to generate auxiliary emotion logits $\mathbf{y}_{\text{emo}}$. A softmax function is then applied to obtain the initial affective estimate:

$$\begin{aligned} \mathbf{y}_{\text{emo}} &= \text{MLP}_{\text{pred}}\big([\mathbf{v}_{\text{base}}; \mathbf{F}_{\text{txt}}^{\text{vis}}]\big), \\ \mathbf{p}_{\text{emo}} &= \text{Softmax}(\mathbf{y}_{\text{emo}}). \end{aligned} \tag{8}$$

where $\mathbf{p}_{\text{emo}} \in \mathbb{R}^C$ denotes the predicted probability distribution over C emotion classes. Instead of collapsing this distribution into a hard categorical prediction, we project the soft distribution back into the embedding space through a trainable dictionary $\mathbf{W}_{\text{emb}} \in \mathbb{R}^{C\times D}$, yielding a continuous affective query $\mathbf{e}_{\text{soft}} \in \mathbb{R}^D$:

$$\mathbf{e}_{\text{soft}} = \mathbf{p}_{\text{emo}}\mathbf{W}_{\text{emb}}. \tag{9}$$

This query $\mathbf{e}_{\text{soft}}$ acts as a semantic filter. By combining it with each local frame descriptor, we derive frame-specific relevance scores s_t through a linear projection followed by temporal normalization:

$$s_t = \frac{\exp\big(\text{Linear}\big([\mathbf{f}_{\text{vis}}^t; \mathbf{e}_{\text{soft}}]\big)\big)}{\sum_{k=1}^{T}\exp\big(\text{Linear}\big([\mathbf{f}_{\text{vis}}^k; \mathbf{e}_{\text{soft}}]\big)\big)}. \tag{10}$$

The context-aware video representation $\mathbf{V}_{\text{agg}} \in \mathbb{R}^D$ is then obtained as

$$\mathbf{V}_{\text{agg}} = \sum_{t=1}^{T} s_t \mathbf{f}_{\text{vis}}^t. \tag{11}$$

By conditioning temporal aggregation on the soft emotion prototype, SETA suppresses irrelevant background content and allocates representational capacity to psychologically meaningful moments. As a result, the temporal aggregation process becomes explicitly hypothesis driven.

3.4. Emotion-Aware Multimodal Fusion

Combining heterogeneous sensory inputs into a unified latent representation remains an ongoing challenge. To address this issue, we adopt a two-step strategy: first aligning

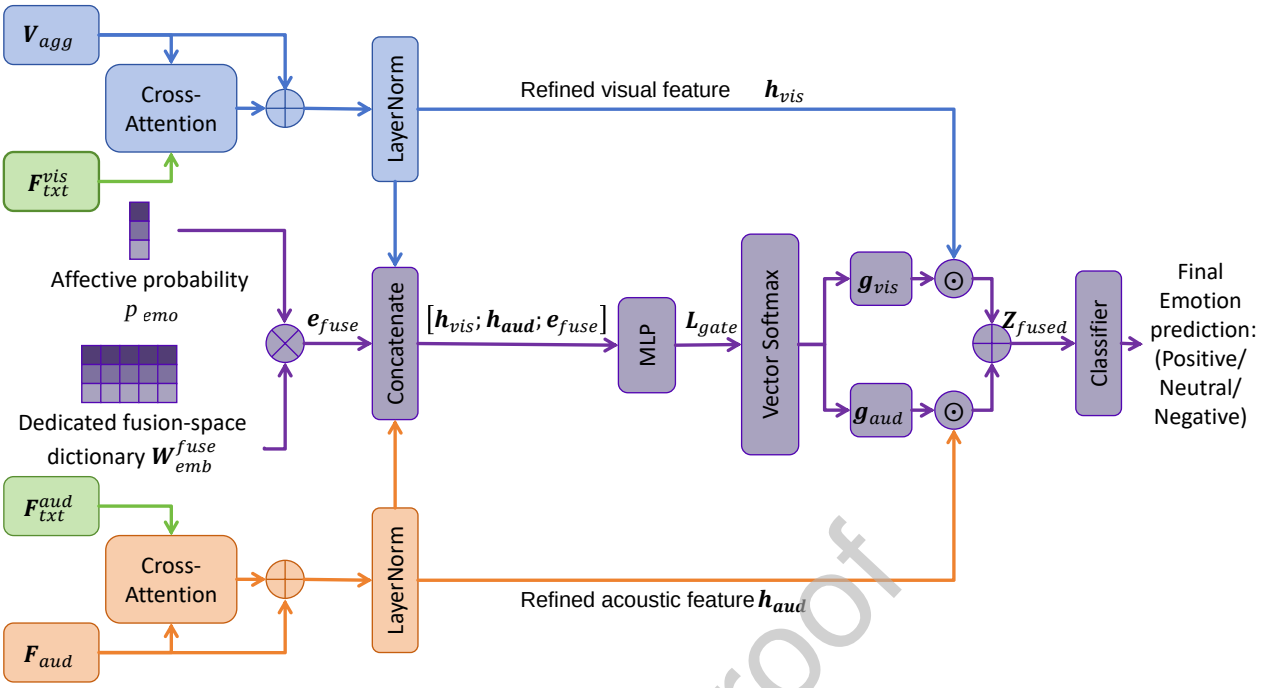


Figure 5: Architecture of the Soft Emotion-Aware Gated Fusion (SEGF) module. Visual and acoustic sensory features are first refined by cross-attention with their corresponding text anchors. The SETA affective probability is projected into a fusion-space query and concatenated with the refined modality features. A vector softmax then produces complementary channel-wise gates $\mathbf{g}_{\text{vis}}$ and $\mathbf{g}_{\text{aud}}$, whose weighted sum forms $\mathbf{Z}_{\text{fused}}$.

modality-specific sensory features with their paired textual guidance, and then applying Soft Emotion-Aware Gated Fusion (SEGF) for dimension-wise adaptive routing.

Description-guided refinement. To inject modality-specific semantic cues, we employ an 8-head cross-attention layer to refine the sensory embeddings. For the visual branch, $\mathbf{V}_{\text{agg}}$ serves as the query, while the visual-semantic anchor $\mathbf{F}_{\text{txt}}^{\text{vis}}$ serves as both the key and the value. After residual addition and layer normalization, the refined visual feature $\mathbf{h}_{\text{vis}} \in \mathbb{R}^D$ is obtained as

$$\mathbf{h}_{\text{vis}} = \text{LayerNorm}\Big(\mathbf{V}_{\text{agg}} + \text{Attention}\big(Q = \mathbf{V}_{\text{agg}},\ K = \mathbf{F}_{\text{txt}}^{\text{vis}},\ V = \mathbf{F}_{\text{txt}}^{\text{vis}}\big)\Big). \tag{12}$$

Similarly, the refined acoustic feature $\mathbf{h}_{\text{aud}} \in \mathbb{R}^D$ is computed as

$$\mathbf{h}_{\text{aud}} = \text{LayerNorm}\Big(\mathbf{F}_{\text{aud}} + \text{Attention}\big(Q = \mathbf{F}_{\text{aud}},\ K = \mathbf{F}_{\text{txt}}^{\text{aud}},\ V = \mathbf{F}_{\text{txt}}^{\text{aud}}\big)\Big). \tag{13}$$

Soft Emotion-Aware Gated Fusion. Unlike conventional flat-weighting strategies that treat all channels equally, SEGF performs fine-grained channel-wise modulation. To maintain consistency with the temporal module, we reuse the emotion probability $\mathbf{p}_{\text{emo}}$ from Section 3.3. Instead of directly reusing the temporal prototype, we project $\mathbf{p}_{\text{emo}}$ through a dedicated fusion dictionary $\mathbf{W}_{\text{emb}}^{\text{fuse}} \in \mathbb{R}^{C\times D}$ to obtain a fusion-specific conditioning vector $\mathbf{e}_{\text{fuse}} \in \mathbb{R}^D$:

$$\mathbf{e}_{\text{fuse}} = \mathbf{p}_{\text{emo}}\mathbf{W}_{\text{emb}}^{\text{fuse}}. \tag{14}$$

The dynamic routing mechanism then evaluates the concatenated joint representation. A routing network MLP_{gate}, consisting of layer normalization, a linear layer, GELU activation, dropout, and an output projection, computes the gating logits:

$$\begin{aligned}&\mathbf{L}_{\text{gate}} = \text{MLP}_{\text{gate}}\big([\mathbf{h}_{\text{vis}}; \mathbf{h}_{\text{aud}}; \mathbf{e}_{\text{fuse}}]\big),\\ &\mathbf{L}_{\text{gate}} \in \mathbb{R}^{2\times D}.\end{aligned} \tag{15}$$

Let $\mathbf{L}_{\text{vis}} \in \mathbb{R}^D$ and $\mathbf{L}_{\text{aud}} \in \mathbb{R}^D$ denote the two rows of $\mathbf{L}_{\text{gate}}$. For each channel k, the modality-specific gating weights are computed using a cross-modality softmax:

$$\begin{aligned}g_{\text{vis}}^{(k)} &= \frac{\exp\big(L_{\text{vis}}^{(k)}\big)}{\exp\big(L_{\text{vis}}^{(k)}\big) + \exp\big(L_{\text{aud}}^{(k)}\big)},\\ g_{\text{aud}}^{(k)} &= \frac{\exp\big(L_{\text{aud}}^{(k)}\big)}{\exp\big(L_{\text{vis}}^{(k)}\big) + \exp\big(L_{\text{aud}}^{(k)}\big)}.\end{aligned} \tag{16}$$

Journal Pre-proof

Journal Pre-proof

The final fused representation $\mathbf{Z}_{\text{fused}} \in \mathbb{R}^D$ is obtained through the Hadamard product:

$$\mathbf{Z}_{\text{fused}} = \mathbf{g}_{\text{vis}} \odot \mathbf{h}_{\text{vis}} + \mathbf{g}_{\text{aud}} \odot \mathbf{h}_{\text{aud}}. \tag{17}$$

This mechanism ensures that the final classifier receives only the most reliable and contextually validated sensory channels.

3.5. Optimization Objective

The overall framework is jointly optimized with a hybrid objective.

Primary classification loss. The main task is to classify the collective emotional state into *Positive*, *Neutral*, or *Negative*. The fused representation $\mathbf{Z}_{\text{fused}}$ is fed into a final linear classifier to produce prediction logits $\mathbf{y}_{\text{pred}}$. A standard cross-entropy loss is then used to supervise the final prediction:

$$\mathcal{L}_{\text{cls}} = -\sum_{c=1}^{C} y_{\text{gt}}^{(c)} \log\left(\text{Softmax}(\mathbf{y}_{\text{pred}})^{(c)}\right). \tag{18}$$

Dual InfoNCE contrastive regularization. To mitigate feature collapse and ensure that the sensory streams internalize the language priors, we introduce a dual InfoNCE loss [25] as a regularizer to align the sensory signals with their corresponding VLM-generated narratives. For a batch of N samples, the unidirectional alignment objective for either modality branch is defined as

$$\mathcal{L}_{\text{InfoNCE}}(\mathbf{x}, \mathbf{y}) = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp\left(\text{sim}(\mathbf{x}_i, \mathbf{y}_i)/\tau\right)}{\sum_{j=1}^{N} \exp\left(\text{sim}(\mathbf{x}_i, \mathbf{y}_j)/\tau\right)}. \tag{19}$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity and τ is the temperature parameter. The total contrastive regularization term is defined as

$$\begin{aligned}\mathcal{L}_{\text{con}} = \frac{1}{2}\Big(&\mathcal{L}_{\text{InfoNCE}}(\mathbf{V}_{\text{agg}}, \mathbf{F}_{\text{txt}}^{\text{vis}}) \\ &+ \mathcal{L}_{\text{InfoNCE}}(\mathbf{F}_{\text{aud}}, \mathbf{F}_{\text{txt}}^{\text{aud}})\Big).\end{aligned} \tag{20}$$

Auxiliary soft-emotion supervision. To ensure that the precursor soft query $\mathbf{e}_{\text{soft}}$ is semantically meaningful, we impose an auxiliary cross-entropy loss on the intermediate logits $\mathbf{y}_{\text{emo}}$ predicted by the SETA module:

$$\mathcal{L}_{\text{emo}} = \text{CE}(\mathbf{y}_{\text{emo}}, \mathbf{y}_{\text{gt}}). \tag{21}$$

Total loss. The final objective is the weighted sum of the above terms:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \lambda_1 \mathcal{L}_{\text{con}} + \lambda_2 \mathcal{L}_{\text{emo}}, \tag{22}$$

where λ_1 and λ_2 control the contributions of the auxiliary regularization terms.

4. Experiments

This section specifies the datasets and evaluation protocol before presenting overall comparisons, ablations, interpretability visualizations, robustness tests, efficiency measurements, case studies, and failure analysis. This order connects each architectural claim to quantitative evidence and then to qualitative behavior.

4.1. Datasets

To evaluate the performance of the proposed framework, we conduct experiments on two representative unconstrained group emotion datasets.

VGAF. VGAF (Video-level Group Affect) [14] is collected from YouTube using semantic queries such as "protest" and "birthday party" and contains 4,183 video clips. The emotion labels are divided into three categories: *Positive*, *Neutral*, and *Negative*. Following the official protocol, the dataset is split into 2,661 training samples, 766 validation samples, and 756 official test samples. Because the ground-truth labels of the official test set are not publicly available, all quantitative results reported in this work are obtained on the validation set, while the training split is used for model optimization.

GECV. GECV (Group-level Emotion on Crowd Videos) [17] provides a challenging benchmark for spontaneous crowd behavior in naturalistic environments. Although the original dataset contains 627 clips with durations ranging from 10 to 20 seconds, privacy restrictions currently limit public access to 408 samples. Similar to VGAF, GECV adopts a three-class taxonomy. Due to the relatively small sample size, we follow a 10-fold cross-validation protocol to obtain statistically reliable performance estimates. Each cross-validation cycle partitions the 408 accessible clips into held-out folds of 40 or 41 clips, with the remaining 367 or 368 clips used for training, so that every accessible sample is evaluated once per cycle.

4.2. Experimental Settings

To facilitate reproducibility, the experimental pipeline is divided into four phases: offline narrative generation, audiovisual preprocessing, end-to-end optimization, and held-out evaluation. The implementation details are described as follows. All experiments are conducted in PyTorch.

For offline semantic extraction, we use the GLM-4.6V-Flash VLM to generate independent visual and acoustic narratives. For representation learning, the visual and textual branches are initialized from CLIP (ViT-B/32) [25], while the auditory branch is built upon CLAP (HTSAT-unfused) [26]. For each video, $T = 20$ frames are uniformly sampled and supplied to the CLIP visual encoder. During data preprocessing, video frames are first resized to 256×256 pixels. During training, random cropping to 224×224 and random horizontal flipping are applied for data augmentation, whereas inference uses a deterministic 224×224 center crop. The RGB channels are normalized using the default

Journal Pre-proof

mean and standard deviation of CLIP. In parallel, raw audio signals are resampled to 48 kHz and truncated by retaining the first 5 seconds from the beginning of each clip to ensure consistent batch dimensions.

The complete trimodal framework is optimized end-to-end using AdamW. The batch size is set to 8, and the initial learning rate is set to 1×10^{-6}. Training is conducted for at most 50 epochs. To alleviate overfitting, early stopping is applied based on validation accuracy, with a patience of 5 epochs. For the joint objective, the contrastive temperature is set to $\tau = 0.07$, while the weighting coefficients are set to $\lambda_1 = 0.5$ for the InfoNCE regularization term and $\lambda_2 = 0.2$ for the auxiliary soft-emotion supervision term. All experiments are conducted on a single NVIDIA RTX Pro 6000 GPU. To reduce statistical bias from a single initialization, the overall evaluation is repeated across 10 independent runs using 10 distinct random seeds. For GECV, each run performs a complete 10-fold cross-validation, resulting in a repeated 10×10-fold evaluation. We report the mean and sample standard deviation across the 10 runs and compute class-wise precision from the classification report of each run. The run with the highest overall accuracy is additionally reported as the best run. Training follows the official class distributions with the standard cross-entropy objective. Evaluation reports overall accuracy, class-wise precision, and run-to-run variation.

For the diagnostic analyses introduced below, the VGAF official validation split is used. Unless otherwise stated, these analyses use the checkpoint associated with the reported best VGAF result.

4.3. Overall Performance

We compare the proposed framework with recent state-of-the-art methods for video-based group emotion recognition. The results on the VGAF and GECV benchmarks are reported in Tables 3 and 4, respectively. For the proposed method, both the mean±SD over 10 runs and the best run are shown, together with class-wise precision. Overall, the classification results support the effectiveness of decoupled linguistic priors and dimension-wise gated fusion on the two benchmarks. These clean-benchmark comparisons are not treated as direct evidence of robustness to unreliable modalities; robustness is evaluated separately through the controlled visual and acoustic corruption experiments in Section 4.6.

Results on VGAF. As shown in Table 3, the proposed method achieves a mean accuracy of 77.69±1.04% across 10 independent runs on the VGAF validation set, while the best run reaches 80.03%.

Recognizing group emotion from complex videos remains challenging. Early audiovisual baselines, such as the action-recognition-based model of Pinto et al. [42] and the static–dynamic fusion model of Sun et al. [18], achieved 62.40% and 70.23% accuracy, respectively. Early cross-modal interaction methods, including those of Petrova et al. [43] and Evtodienko [21], reported 52.36% and 60.37%. More recent graph-based and transformer-based methods, including Savchenko et al. [44], Wang et al. [13], and Huang and Xu [16], report accuracies in the 66–70% range. Augusma et al. [33] achieved 78.72%, while the best run of the proposed framework reaches 80.03%, a difference of 1.31 percentage points.

Slogrove and van der Haar [32] reported 91% using pose estimation and an LSTM on VGAF. However, their reported confusion matrix covers 226 clips rather than the official 766-video validation partition, the exact split construction is not fully specified, and no public implementation is available. We therefore cite this result for completeness but do not place it in Table 3, where all entries are intended to follow a directly comparable validation protocol.

The per-run classification reports provide a more detailed view of class behavior. The mean precision values are 82.37±2.74%, 71.50±2.12%, and 81.95±3.65% for the positive, neutral, and negative classes, respectively. The neutral class remains the most difficult in terms of precision, whereas the positive and negative classes show stronger discrimination. In the best run, the corresponding precision values are 86.02%, 72.64%, and 84.81%.

Results on GECV. The GECV dataset contains substantial real-world challenges, such as camera motion, illumination variation, and strong acoustic interference. As shown in Table 4, the proposed method achieves a mean accuracy of 97.87±0.47% across 10 repeated cross-validation runs, and the best run reaches 98.52%.

The 10-run mean surpasses the spatial-temporal graph model of Wang et al. [13] by 4.97 percentage points and the multimodal spatiotemporal semi-supervised transformer of Huang and Xu [16] by 11.00 points; the best run increases these differences to 5.62 and 11.65 points, respectively. Quach et al. [17], who introduced GECV, reported 70.97% with TNVPF under a 90/10 split on the original 627 videos. The proposed evaluation uses the currently accessible 408 clips and repeated 10-fold cross-validation.

The class-wise results are consistently high across repeated evaluations. The mean precision values are 98.34±0.68%, 97.05±1.05%, and 98.49±1.15% for positive, neutral, and negative samples, respectively. In the best run, the corresponding values are 98.61%, 97.26%, and 100.00%. The small variation in overall accuracy and class-wise precision indicates stable classification under repeated cross-validation. These clean-benchmark results characterize predictive performance but do not, by themselves, demonstrate robustness to unreliable modalities. Robustness is assessed separately in Section 4.6 through controlled visual and acoustic corruption experiments and matched clean-versus-corrupted comparisons.

4.4. Ablation Study

To quantitatively assess the contribution of each architectural component, we conduct a comprehensive ablation study on the VGAF dataset. The framework is analyzed along three main dimensions: the effect of decoupled semantic priors, the benefits of hypothesis-guided temporal

Journal Pre-proof

Table 3
Comparison of classification performance on the VGAF dataset. Bold indicates the best value for each metric when Ours (best) is compared with the published baselines; Ours (mean) is reported for stability and is excluded from the boldface comparison. "–" indicates an unreported value. Unit: %.

Methods	Accuracy	Precision		
		Pos.	Neu.	Neg.
Pinto et al. [42]	62.40	69.20	52.50	66.30
Sun et al. [18]	70.23	–	–	–
Petrova et al. [43]	52.36	60.00	40.00	80.00
Evtodienko [21]	60.37	51.00	**73.00**	56.00
Savchenko et al. [44]	70.23	73.20	64.30	70.70
Augusma et al. [33]	78.72	–	–	–
Wang et al. [13]	70.23	–	–	–
Huang and Xu [16]	66.42	60.41	70.26	71.32
Ours (best)	**80.03**	**86.02**	72.64	**84.81**
Ours (mean)	77.69±1.04	82.37±2.74	71.50±2.12	81.95±3.65

Table 4
Comparison of classification performance on the GECV dataset. Bold indicates the best value for each metric when Ours (best) is compared with the published baselines; Ours (mean) is reported for stability and is excluded from the boldface comparison. †Quach et al. used the original 627-video dataset with a 90/10 train–test split; the proposed results use the currently accessible 408 clips and repeated 10-fold cross-validation. "–" indicates an unreported value. Unit: %.

Methods	Accuracy	Precision		
		Pos.	Neu.	Neg.
Quach et al. [17]†	70.97	–	–	–
Wang et al. [13]	92.90	96.65	90.56	89.75
Huang and Xu [16]	86.87	90.31	82.43	80.17
Ours (best)	**98.52**	**98.61**	**97.26**	**100.00**
Ours (mean)	97.87±0.47	98.34±0.68	97.05±1.05	98.49±1.15

Table 5
Modality-level ablation on the VGAF dataset. The best value for each metric is shown in bold. Unit: %.

Modalities	Accuracy	Precision		
		Pos.	Neu.	Neg.
Visual	70.63	73.94	**75.00**	61.84
Audio	62.92	67.48	60.91	60.53
Visual + Audio	71.41	74.92	65.31	75.84
Visual + Visual Text	74.15	80.26	66.89	75.90
Audio + Audio Text	68.54	**87.85**	63.42	61.38
Visual + Visual Text + Audio	77.42	82.27	72.20	77.91
Audio + Audio Text + Visual	71.93	75.69	70.31	67.57
Ours (full)	**80.03**	86.02	72.64	**84.81**

Journal Pre-proof

and channel-wise routing, and the contribution of the hybrid training objective.

Effect of decoupled semantics. We first verify the central assumption of our framework: explicitly injecting modality-specific semantic priors yields better representations than relying solely on raw sensory signals. Table 5 reports the classification performance under different combinations of raw modalities and their corresponding linguistic priors.

An initial comparison of the unimodal branches indicates that the visual stream has stronger discriminative power than the audio stream, achieving 70.63% and 62.92% accuracy, respectively. Simply combining the two raw modalities yields an accuracy of 71.41%, suggesting that the direct fusion of low-level sensory features provides only limited improvement. In contrast, the introduction of decoupled linguistic anchors leads to a clear performance gain. Specifically, combining visual data with its dedicated textual narrative improves accuracy to 74.15%, while injecting the audio-specific semantic prior increases the acoustic baseline to 68.54%. These results confirm that theory-guided parallel VLM prompting provides clean and informative conceptual supervision.

The asymmetric combinations further reveal the relative importance of the two modalities. A configuration centered on visual semantics (Visual + Visual Text + Audio) achieves 77.42%, significantly outperforming the audio-centered counterpart (Audio + Audio Text + Visual), which reaches 71.93%. This gap suggests that visual behavior is the

Journal Pre-proof

Table 6
Module-level ablation on the VGAF dataset. The best value for each metric is shown in bold. Unit: %.

Temporal	Fusion	Accuracy	Precision		
			Pos.	Neu.	Neg.
Pooling	Concat	73.50	80.90	64.67	81.10
SETA	Concat	76.24	79.11	69.70	83.01
Pooling	SEGF	75.85	83.40	67.17	81.07
SETA	SEGF	**80.03**	**86.02**	**72.64**	**84.81**

Table 7
Loss-level ablation on the VGAF dataset. The best value for each metric is shown in bold. Unit: %.

Loss configuration	Accuracy	Precision		
		Pos.	Neu.	Neg.
$\mathcal{L}_{\text{cls}}$	74.93	79.55	68.28	79.17
$\mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{con}}$	76.24	79.88	69.16	83.70
$\mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{emo}}$	75.46	77.78	70.61	80.29
Ours (full objective)	**80.03**	**86.02**	**72.64**	**84.81**

dominant cue for group emotion recognition, whereas audio mainly provides auxiliary contextual information. Under the fully decoupled setting, the proposed framework achieves the best overall performance, reaching 80.03% accuracy and 86.02% precision for positive samples. Preserving the independence of visual and acoustic reasoning before final fusion effectively mitigates narrative interference, thereby validating the core design principle of the proposed architecture.

Module-level comparison. To analyze the specific contributions of SETA and SEGF, we compare them with conventional aggregation strategies that do not incorporate affective context. The results are reported in Table 6.

Table 6 shows that a basic architecture using average pooling and direct feature concatenation achieves an accuracy of 73.50%. This result reflects the limitations of class-agnostic aggregation, which smooths transient affective peaks and accepts all feature dimensions without explicit selection. Replacing average pooling with SETA, while retaining simple concatenation, improves the accuracy to 76.24% and raises the precision for the negative class to 83.01%. This improvement indicates that using an initial soft-emotion hypothesis to guide temporal scanning helps identify emotionally salient moments that would otherwise be diluted by naive averaging.

By contrast, using average pooling together with SEGF achieves 75.85% accuracy. Even when temporally smoothed features are used as input, SEGF still acts as an effective channel-wise filter that dynamically suppresses unreliable sensory information. Finally, combining SETA and SEGF yields the best performance, reaching 80.03% accuracy. In this joint configuration, SETA provides temporally refined features, enabling SEGF to perform more accurate and context-aware channel selection.

Role of the hybrid training objective. We further evaluate the necessity of the three-part training objective. In addition to the primary classification loss $\mathcal{L}_{\text{cls}}$, the optimization of the network is regularized by the contrastive term $\mathcal{L}_{\text{con}}$ and the auxiliary affective supervision term $\mathcal{L}_{\text{emo}}$. To quantify their individual and combined contributions, we evaluate these objectives in isolation and in combination, as shown in Table 7.

Table 7 shows that optimizing the model with $\mathcal{L}_{\text{cls}}$ alone is insufficient, yielding only 74.93% accuracy and relatively weak performance on the neutral class (68.28%). Adding the auxiliary supervision term $\mathcal{L}_{\text{emo}}$ improves the accuracy to 75.46%, indicating that encouraging SETA to predict an early affective distribution helps construct a more informative temporal attention map. More importantly, incorporating the contrastive alignment term $\mathcal{L}_{\text{con}}$ raises the accuracy to 76.24%, supporting the necessity of active cross-modal alignment. By aligning sensory features with their corresponding VLM-generated textual anchors, $\mathcal{L}_{\text{con}}$ encourages the encoders to learn more robust and semantically structured representations, thereby reducing confusion between visually similar but acoustically distinct scenarios. The full hybrid objective achieves the best overall performance at 80.03% accuracy, demonstrating that the three loss terms are complementary.

4.5. Interpretability Analysis

This subsection examines whether the internal behavior of SETA, SEGF, and the soft emotional prototype is consistent with the intended design. The analysis complements the quantitative ablations by visualizing temporal attention, class-wise gating behavior, and uncertainty in the prototype distribution.

SETA temporal attention. Figure 6 visualizes the normalized temporal attention weights learned by SETA for three illustrative VGAF validation samples, while Fig. 7 shows the corresponding sampled frames. The attention distributions are clearly non-uniform: the maximum weights are approximately 0.065, 0.068, and 0.072 for the illustrated positive, neutral, and negative samples, respectively. For the positive sample, high weights cluster around later frames; the neutral sample assigns its largest weights to a small number of early and middle frames; and the negative sample concentrates attention near the final segment. The highlighted frames therefore provide qualitative evidence that SETA selectively emphasizes temporally localized, emotionally salient evidence instead of treating all frames equally.

SEGF gating behavior. Figure 8 summarizes the class-wise SEGF gate values. The mean visual gate weights are 0.5123, 0.5109, and 0.5091 for positive, neutral, and negative samples, respectively, with the complementary audio weights given by $g_{\text{aud}} = 1 - g_{\text{vis}}$. The values remain close to the equal-weighting point of 0.5, indicating that the model does not collapse to a single modality. The small category-dependent shift should therefore be interpreted as a mild tendency rather than strong modality dominance: SEGF

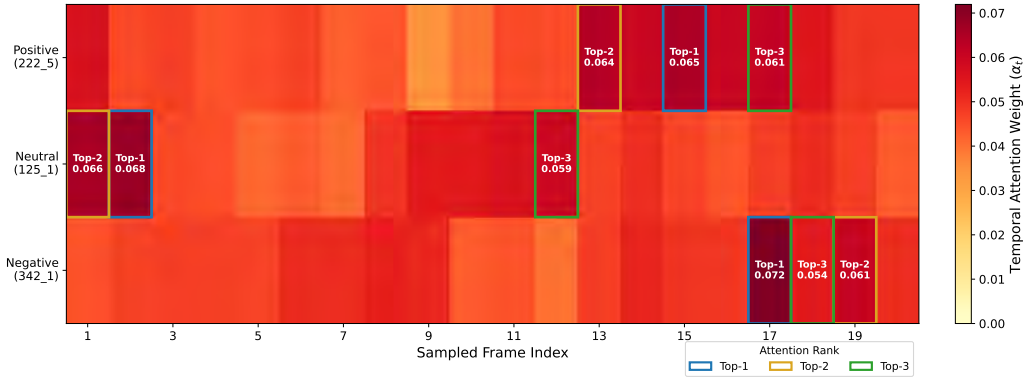


Figure 6: SETA temporal attention weights for illustrative VGAF validation samples from the three emotion classes. Each row denotes one video and the columns denote 20 sampled frames; colored boxes indicate the three largest temporal attention weights.

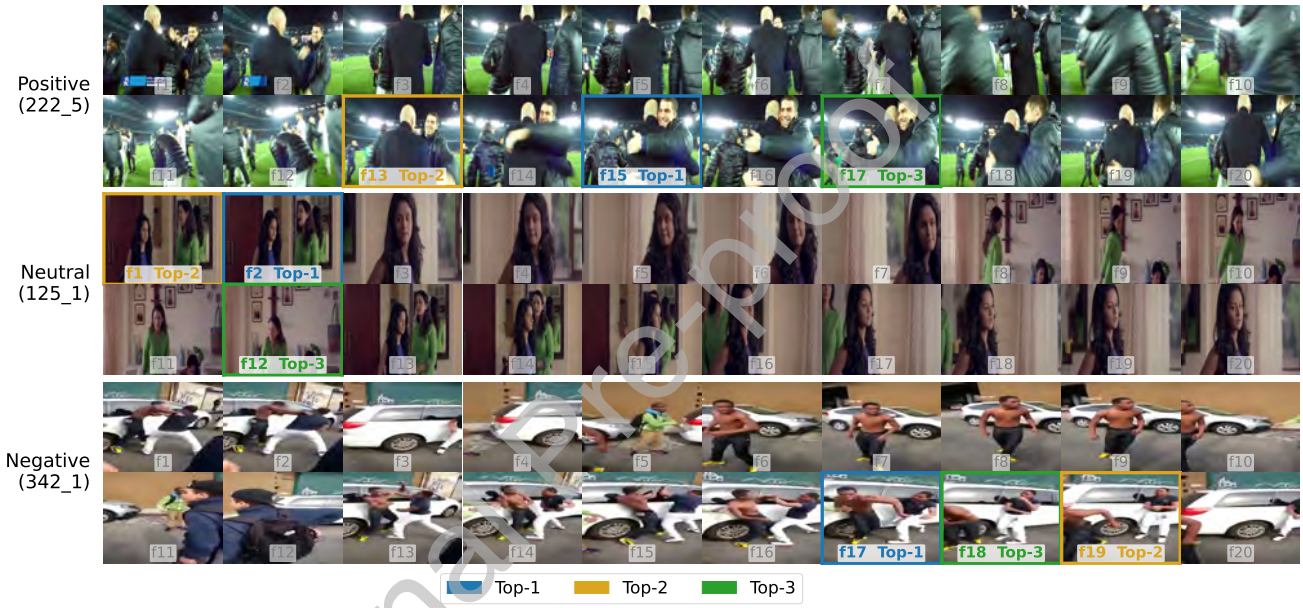


Figure 7: Corresponding video frames for the samples in Fig. 6. The colored borders mark the top-three frames ranked by SETA attention weight, allowing the learned temporal emphasis to be inspected directly against the visual content.

maintains approximately balanced multimodal routing while allowing sample-dependent channel modulation.

Interpretation of soft emotional prototypes. We further analyze the entropy of the soft emotion distribution $\mathbf{p}_{\text{emo}}$ over all 766 VGAF validation samples; the maximum possible entropy for three classes is $\ln 3 = 1.099$. For diagnostic analysis, we define low-entropy samples as $H < 0.50$ and high-entropy samples as $H > 0.85$. These cutoffs correspond to approximately 46% and 77% of the maximum entropy, respectively, and create clearly separated concentrated and diffuse prototype regimes while retaining sufficiently large groups ($n = 209$ and $n = 239$) for class-wise analysis. The thresholds are used only to interpret model behavior and are not optimized as decision rules. As shown in Fig. 9 and Table 8, the low-entropy group attains 95.22% accuracy, whereas the high-entropy group attains 69.04%. The low-entropy group also achieves precision values of 94.85/93.75/97.92% for Positive/Neutral/Negative samples, compared with 75.68/62.50/75.56% for the high-entropy group. Low entropy therefore corresponds to a concentrated prototype hypothesis, as expected for coherent and intense collective affect, whereas high entropy reflects ambiguous or mixed evidence. Importantly, the soft distribution preserves this uncertainty instead of prematurely collapsing the affective state into a hard one-hot label.

4.6. Robustness under Sensory Corruption

To assess behavior under degraded sensory inputs, we inject representative visual and acoustic corruptions into the VGAF validation set.

Visual corruption. Figure 10 reports results for random frame dropping, random spatial occlusion, and Gaussian blur. Accuracy decreases from 80.03% to 75.59% when all visual frames are dropped. Spatial occlusion is more

Journal Pre-proof

Journal Pre-proof

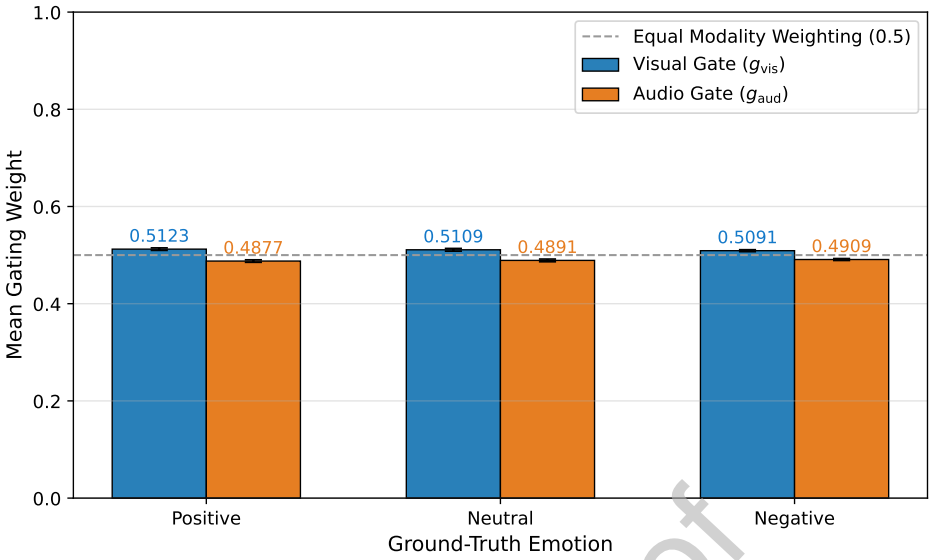


Figure 8: Class-wise mean SEGF visual and audio gate weights on the VGAF validation set. The dashed line denotes equal modality weighting.

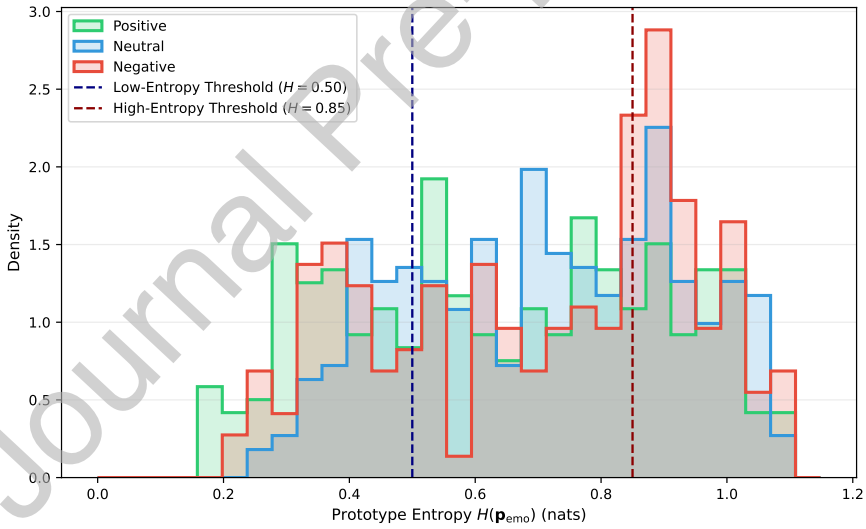


Figure 9: Class-wise distribution of soft-prototype entropy on the VGAF validation set. The dashed lines denote the diagnostic low-entropy threshold $H = 0.50$ and high-entropy threshold $H = 0.85$.

Table 8
Diagnostic performance of the low- and high-entropy groups on the VGAF validation set. The best value for each metric is shown in bold. Unit: %.

Entropy group	Accuracy	Precision		
		Pos.	Neu.	Neg.
Low ($H < 0.50$, $n = 209$)	**95.22**	**94.85**	**93.75**	**97.92**
High ($H > 0.85$, $n = 239$)	69.04	75.68	62.50	75.56

Journal Pre-proof

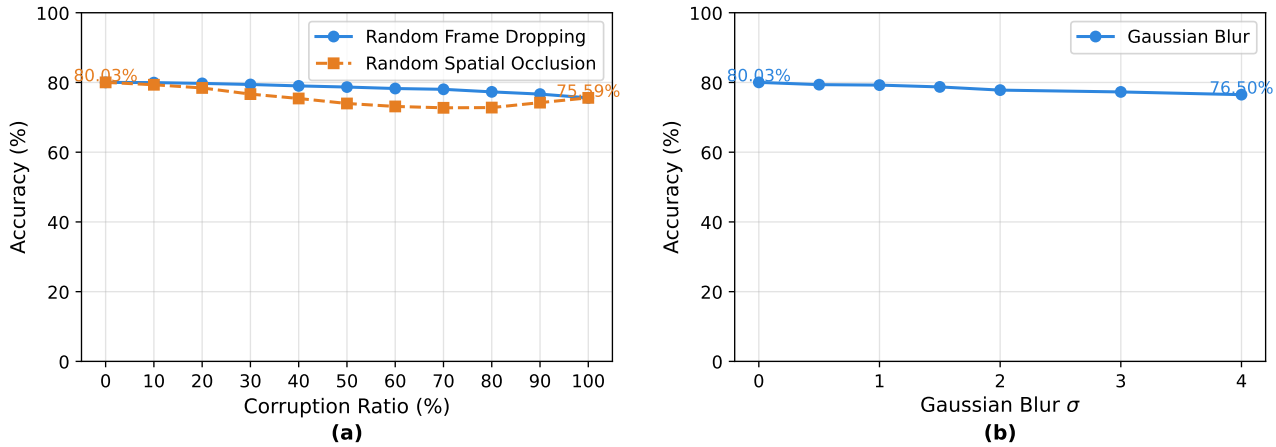


Figure 10: VGAF validation accuracy under visual corruption. (a) Random frame dropping and random spatial occlusion. (b) Gaussian blur.

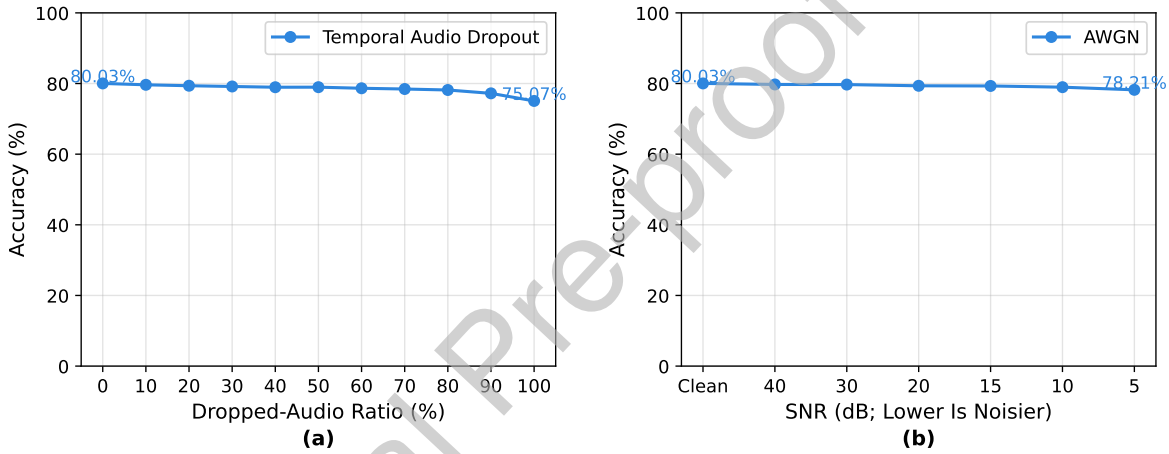


Figure 11: VGAF validation accuracy under acoustic corruption. (a) Temporal audio dropout. (b) Additive white Gaussian noise. The first point in each panel is the clean-input condition.

disruptive: accuracy falls to 74.22% at a 90% occluded-area ratio. Interestingly, when the visual signal is completely removed, accuracy recovers to 75.59%, suggesting that partially corrupted visual evidence can be more harmful than a fully absent stream because the remaining noisy features still enter multimodal fusion. Gaussian blur produces a smoother degradation, reaching 76.50% at $\sigma = 4$. Overall, the curves show graceful performance degradation rather than abrupt failure.

Acoustic corruption. Figure 11(a) reports the temporal audio-dropout experiment. The model retains 78.36% accuracy even when 80% of the acoustic signal is removed and reaches 75.07% when the audio is completely suppressed. Figure 11(b) reports additive white Gaussian noise, under which accuracy remains 78.21% at 5 dB SNR. These results indicate that the model can exploit complementary visual and semantic information when the acoustic stream is degraded.

The corruption curves and the module ablation in Table 6 provide complementary evidence for robust multimodal recognition. Adding SEGF raises accuracy from 73.50% to 75.85% with average pooling, and from 76.24% to 80.03% when combined with SETA, supporting the contribution of channel-wise gated routing.

4.7. Efficiency Analysis

We next quantify the computational cost of the proposed framework. Figure 12 shows that the complete model contains 308.59 M parameters. The task-specific modules comprising SETA, the two cross-attention blocks, SEGF, and the classifier account for only 3.823 M parameters, approximately 1.2% of the total. The profiled computation is 93.59 GFLOPs per sample for the batch-size-one, 20-frame profiling configuration. The CLIP visual encoder dominates the computation at 87.32 GFLOPs, followed by the CLAP audio encoder at 5.69 GFLOPs; the CLIP and CLAP text encoders contribute 0.15 and 0.43 GFLOPs, respectively,

Journal Pre-proof

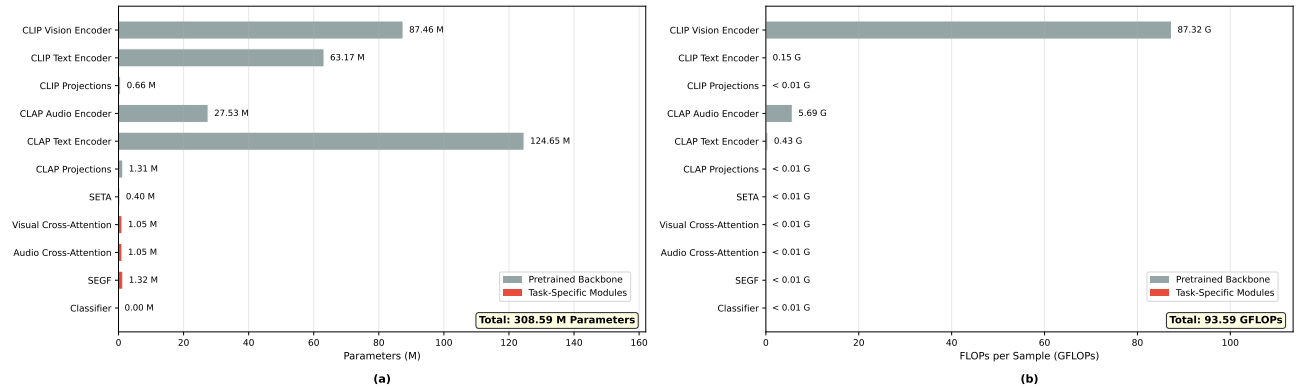


Figure 12: Efficiency profile of the proposed framework. (a) Parameter breakdown by component. (b) FLOP breakdown per 20-frame sample.

Table 9
Model efficiency on a single NVIDIA RTX Pro 6000 GPU. Latency is measured with batch size 1.

Metric	Value
Total parameters	308.59 M
Task-specific module parameters	3.823 M (1.2%)
FLOPs per sample (20 frames)	93.59 G
Inference latency (mean±SD)	59.8±7.2 ms
Median latency (P50)	60.1 ms
P95 latency	69.5 ms

while each proposed task-specific module contributes less than 0.01 GFLOPs.

The VLM narrative generation is performed as offline preprocessing in the present experimental protocol and is therefore excluded from the model-only inference latency in Table 9. The measured visual and acoustic narrative-generation stages require approximately 12.6 and 12.4 seconds per sample, respectively, including API/network overhead (two successful measurements for each branch). The recognition network requires 59.8±7.2 ms per sample, with a median latency of 60.1 ms and a P95 latency of 69.5 ms. Thus, although the proposed task-specific modules add little computational overhead, the overall system still depends on large pretrained encoders and offline VLM access. Direct edge deployment would require further compression, such as parameter-efficient adaptation, quantization, distillation, or replacement with smaller encoders.

4.8. Case Study: Decoupled versus Shared Narratives

To close the loop between the motivation and the proposed modality-decoupling strategy, we perform a paired evaluation on all 766 VGAF validation samples. In the decoupled condition, the original psychology-informed visual and acoustic descriptions are supplied to their corresponding text branches. For the shared-description condition, we merge the visual and acoustic prompt instructions and submit the audiovisual content to GLM-4.6V-Flash to generate one unified description. This same text is then used simultaneously as the visual and acoustic description input, while the raw sensory inputs, checkpoint, and remaining inference settings are kept unchanged.

As shown in Table 10, the decoupled narratives achieve 80.03% accuracy, with class-wise precision of 82.62%, 78.49%, and 78.36% for the Positive, Neutral, and Negative classes, respectively. The shared description obtains 76.50% accuracy and class-wise precision of 80.34%, 74.35%, and 73.81%, corresponding to decreases of 3.53 percentage points in accuracy and 2.28, 4.14, and 4.55 points in the three class-wise precision values. This controlled comparison indicates that a single blended description is vulnerable when visual and acoustic evidence must remain distinguishable.

Figure 13 reports the prediction probabilities and corresponding qualitative evidence for sample 209_15, whose ground-truth label is *Negative*. With the original modality-specific narratives, the model correctly predicts *Negative* with probabilities $P(\text{Positive}) = 0.1870$, $P(\text{Neutral}) = 0.0066$, and $P(\text{Negative}) = 0.8064$. When the unified description is duplicated across both text branches, the prediction flips to *Positive*, with probabilities 0.5503, 0.0080, and 0.4417, respectively. The value of the decoupled narratives therefore lies not merely in adding more words, but in retaining branch-specific semantic structures that condition visual and acoustic evidence differently. This result directly links the semantic-entanglement motivation to the proposed modality-decoupled design and its quantitative validation.

4.9. Failure Case Analysis

To identify where the proposed mechanisms remain unreliable, we inspect the 153 misclassified samples produced by the best VGAF checkpoint. Figure 14 presents four complementary error patterns rather than only visually obvious mistakes: a high-confidence error, a high-entropy ambiguous case, a low-confidence decision boundary, and a low-entropy but confidently wrong prototype. The red and orange borders indicate the two frames receiving the highest SETA weights, while the reported probability, entropy, and mean gate values expose the internal state of the model.

Journal Pre-proof

Table 10
Accuracy and class-wise precision under decoupled and shared narrative settings on the complete 766-sample VGAF validation set. Bold indicates the higher value for each metric. Unit: %.

Narrative setting	Accuracy	Precision		
		Pos.	Neu.	Neg.
Decoupled narratives	**80.03**	**82.62**	**78.49**	**78.36**
Shared description	76.50	80.34	74.35	73.81

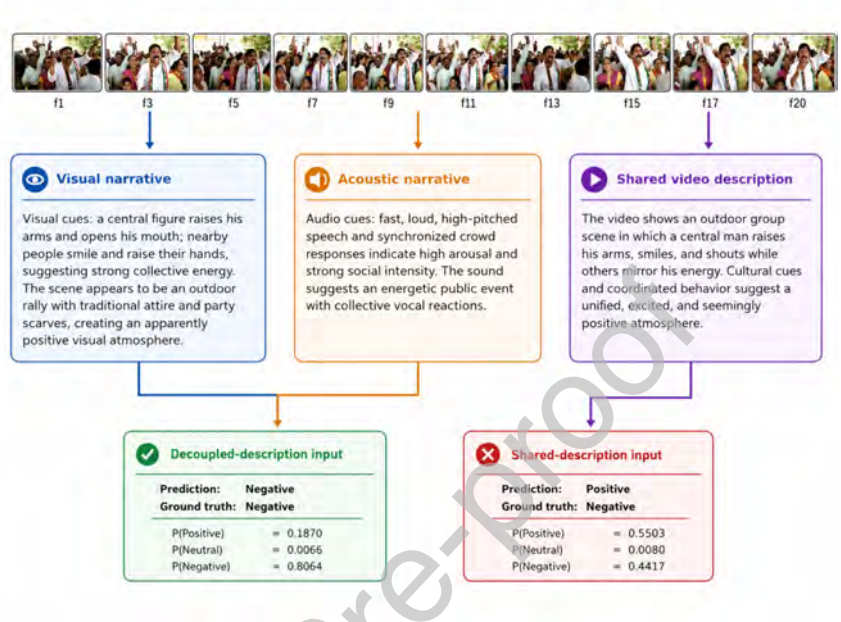


Figure 13: Qualitative visualization and prediction probabilities for sample 209_15 from the VGAF validation set. Replacing the modality-specific narratives with one shared description changes the correct *Negative* prediction to *Positive*.

In Fig. 14(a), sample 284_5 is labeled *Positive* but is predicted as *Neutral* with 0.980 confidence; both generated narratives emphasize calm, neutral interaction, suggesting that subtle low-arousal positive cues can be overwhelmed by semantically coherent neutral descriptions. Sample 143_5 in Fig. 14(b) has an almost uniform soft prototype (normalized entropy 0.992) and a low maximum class probability of 0.426, which is consistent with genuinely ambiguous evidence. In Fig. 14(c), the visual narrative describes a neutral formal ceremony while the acoustic narrative emphasizes a warm celebratory atmosphere; the resulting near-boundary prediction illustrates unresolved cross-modal conflict. Figure 14(d) is more challenging: the model forms a low-entropy negative prototype and a 0.977-confidence *Negative* decision for a sample annotated as *Neutral*. The visually coherent crowd scene therefore produces a clear but incorrect affective hypothesis, indicating that contextually plausible cues can still reinforce a dataset-label mismatch or a culturally dependent interpretation.

Across all four examples, the mean visual and audio gates remain close to 0.5. SEGF can suppress locally noisy channels, but balanced gates cannot correct a situation in which both streams are individually coherent yet jointly biased toward the wrong semantic interpretation. These observations delimit the model's practical reliability and motivate future work on calibration, cross-cultural annotation consistency, and uncertainty-aware abstention.

4.10. Zero-Shot Recognition with Vision-Language Models

To compare the proposed task-specific framework with direct VLM classification, we construct a zero-shot baseline using GLM-4.6V-Flash.

To ensure a consistent evaluation protocol, the generative model is provided with the same input information as the proposed framework, including raw video frames and the precomputed modality-decoupled textual narratives. The prompting strategy follows the psychology-informed instructions introduced in Section 3.1, with the additional requirement that the model return its three-way prediction (*Positive*, *Neutral*, or *Negative*) in JSON format.

The results in Tables 11 and 12 reveal a substantial performance gap in the zero-shot setting. Although the standalone VLM demonstrates a basic understanding of crowd dynamics, its classification performance remains well below that of the proposed framework. On VGAF, the best run of our method outperforms the direct zero-shot VLM baseline by 11.14 percentage points, improving the accuracy from

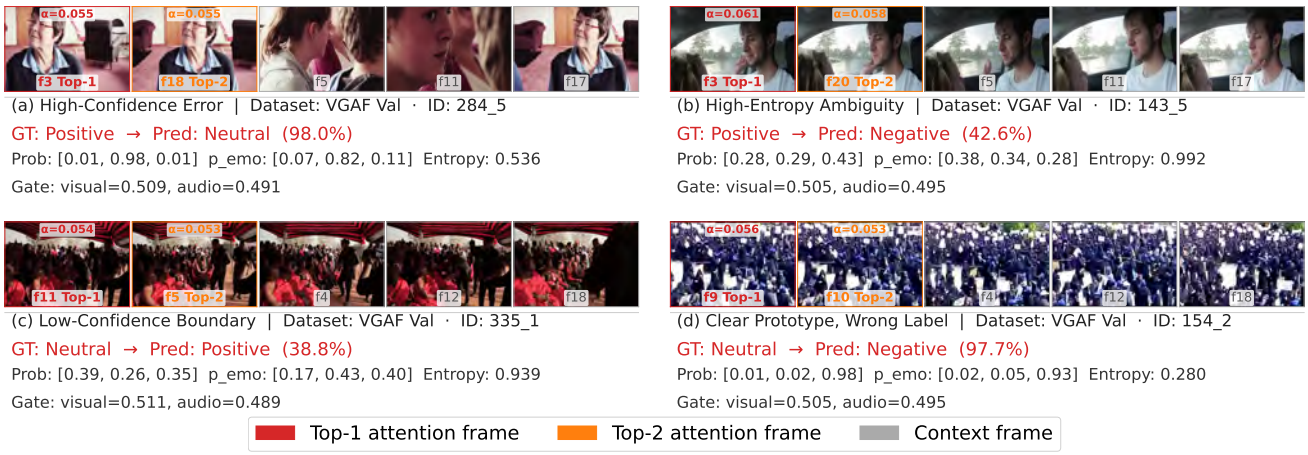


Figure 14: Representative failure cases from the VGAF validation set using the best checkpoint. (a) High-confidence error: a subtle positive interaction is described and classified as neutral. (b) High-entropy ambiguity: the soft prototype is nearly uniform and the final confidence is low. (c) Low-confidence boundary: visual and acoustic narratives support conflicting neutral and positive interpretations. (d) Clear prototype but wrong label: a visually coherent crowd scene produces a confident negative prediction for a neutral annotation. Red and orange borders denote the top-1 and top-2 SETA frames; α is the temporal attention weight, entropy is $H(\mathbf{p}_{\text{emo}})/\log 3$, Prob and $\mathbf{p}_{\text{emo}}$ list Positive/Neutral/Negative probabilities, and Gate(v/a) reports the mean SEGF visual/audio weights.

Table 11
Zero-shot affective classification results of GLM-4.6V-Flash on the VGAF validation set. The best value for each metric is shown in bold. Unit: %.

Methods	Accuracy	Precision		
		Pos.	Neu.	Neg.
GLM-4.6V-Flash (zero-shot)	68.89	65.31	69.57	80.34
Ours (best run)	**80.03**	**86.02**	**72.64**	**84.81**

Table 12
Zero-shot affective classification results of GLM-4.6V-Flash on the GECV benchmark. The best value for each metric is shown in bold. Unit: %.

Methods	Accuracy	Precision		
		Pos.	Neu.	Neg.
GLM-4.6V-Flash (zero-shot)	81.77	77.87	93.94	83.33
Ours (best run)	**98.52**	**98.61**	**97.26**	**100.00**

68.89% to 80.03%. The gap becomes even larger on GECV, where the best run achieves a 16.75-point improvement over the zero-shot baseline (98.52% versus 81.77%).

The accuracy gaps in Tables 11 and 12 show that the complete proposed framework outperforms direct zero-shot classification with GLM-4.6V-Flash under the evaluated protocols. This comparison does not isolate the effects of temporal smoothing, channel-wise noise suppression, or any individual module. Those mechanism-specific claims are supported separately by the SETA and SEGF ablations, the attention and gate analyses, and the controlled corruption experiments in Sections 4.4–4.6. The zero-shot results should therefore be interpreted as evidence for the advantage of the complete task-specific framework over the evaluated zero-shot VLM baseline, rather than as a causal test of SETA or SEGF.

4.11. Discussion, Practical Scope, and Limitations

This subsection critically interprets the results relative to prior work and states the practical scope of the proposed method. On VGAF, the best run is 1.31 percentage points above the 78.72% result of Augusma et al. [33], whereas the 10-run mean of 77.69% is 1.03 points lower. This combination indicates competitive average performance together with a strong best-run result, and the class-wise precision, repeated-run variation, temporal attention, gating, corruption, and failure analyses provide a broader characterization than accuracy alone. On GECV, Table 4 reports each method together with its evaluation protocol.

Relative to pose/LSTM, graph, transformer, and conventional fusion schemes, the main advantages of the framework are explicit visual–acoustic narrative separation, interpretable emotion-conditioned frame selection, and feature-channel-level routing. The framework also presents several practical limitations. Its computational pipeline depends on large pretrained CLIP/CLAP encoders and offline VLM

Journal Pre-proof

Journal Pre-proof

access, so edge deployment requires smaller encoders, distillation, quantization, or parameter-efficient adaptation. The available GER datasets remain modest in size and contain context-dependent labels, motivating further study of cross-dataset generalization, calibration, cultural fairness, and privacy. Established psychological theories serve as design priors for prompt construction and feature interpretation within the computational GER framework.

The intended application is analyst-in-the-loop indexing and retrieval of crowd videos, multimedia event analysis, and retrospective monitoring in which model outputs can prioritize clips for human review. The failure cases in Section 4.9 show why the current system should not be used as an autonomous safety-critical decision maker. Future work will investigate uncertainty-aware abstention, robust narrative generation under degraded inputs, smaller or distilled encoders, parameter-efficient adaptation, and external validation across events and cultural contexts.

5. Conclusion

In this work, we address cross-modal semantic leakage, temporal smoothing, and sensory noise in collective emotion analysis through a modality-decoupled language-guided framework. The principal novelty is the integration of three mutually reinforcing ideas: separate visual and acoustic VLM narratives, a soft emotion prototype that actively queries temporally salient frames through SETA, and the reuse of that prototype to condition channel-wise multimodal routing through SEGF. Dual InfoNCE alignment connects the decoupled semantic anchors to their corresponding sensory streams. Across 10 independent runs, the proposed method achieves mean accuracies of 77.69±1.04% and 97.87±0.47% on VGAF and GECV, respectively, with best-run accuracies of 80.03% and 98.52%. The interpretability, robustness, efficiency, narrative-decoupling, and failure-case analyses support the intended behavior while also revealing limitations: the current system depends on offline VLM descriptions and large pretrained encoders, and context-dependent or culturally ambiguous samples can still be confidently misclassified. Future work will therefore focus on uncertainty-aware abstention, robust narrative generation under degraded inputs, cross-dataset and cross-cultural validation, and parameter-efficient compression for resource-constrained deployment.

Credit Author Statement

Hongxi Yi: Methodology, Investigation, Conceptualization, Data curation, Software, Visualization, Writing original draft; Yufei Xie: Investigation, Formal analysis, Conceptualization, Data curation, Software; Dong Zhang: Methodology, Investigation, Conceptualization, Writing – review & editing, Funding acquisition; Wenxing Liu: Investigation, Conceptualization; Ming Li: Investigation, Conceptualization Dah-Jye Lee: Writing – review & editing.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] E. Hatfield, J. T. Cacioppo, and R. L. Rapson. Emotional contagion. *Current Directions in Psychological Science*, 2(3):96–100, 1993.

[2] S. G. Barsade. The ripple effect: Emotional contagion and its influence on group behavior. *Administrative Science Quarterly*, 47 (4):644–675, 2002.

[3] K. Fujii, D. Sugimura, and T. Hamamoto. Hierarchical group-level emotion recognition in the wild. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–5. IEEE, 2019.

[4] K. Fujii, D. Sugimura, and T. Hamamoto. Hierarchical group-level emotion recognition. *IEEE Transactions on Multimedia*, 23:3892–3906, 2020.

[5] A. S. Khan, Z. Li, J. Cai, et al. Regional attention networks with context-aware fusion for group emotion recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1150–1159, 2021.

[6] B. Nagarajan and V. R. M. Oruganti. Group emotion recognition in adverse face detection. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–5. IEEE, 2019. doi: 10.1109/FG.2019.8756553.

[7] X. Wang, D. Zhang, and D. J. Lee. Implementing the affective mechanism for group emotion recognition with a new graph convolutional network architecture. *IEEE Transactions on Affective Computing*, 15 (3):1104–1115, 2023.

[8] X. Guo, L. Polania, B. Zhu, et al. Graph neural networks for image understanding based on multiple cues: Group emotion recognition and event recognition as use cases. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2921–2930, 2020.

[9] Y. Dai, J. Li, Y. Li, et al. Multi-modal graph context extraction and consensus-aware learning for emotion recognition in conversation. *Knowledge-Based Systems*, 298:111954, 2024.

[10] C. Li, W. Huang, X. Yu, et al. Group emotion recognition using fuzzy emotional hierarchical shared space. *IEEE Transactions on Fuzzy Systems*, 2025.

[11] Q. Zhu, Q. Mao, J. Zhang, et al. Towards a robust group-level emotion recognition via uncertainty-aware learning. *IEEE Transactions on Affective Computing*, 2025.

[12] D. Tran, H. Wang, L. Torresani, et al. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.

[13] X. Wang, T. Chen, and D. Zhang. A spatial-temporal graph convolutional network for video-based group emotion recognition. In *International Conference on Pattern Recognition*, pages 339–354, Cham, 2024. Springer Nature Switzerland.

[14] G. Sharma, A. Dhall, and J. Cai. Audio-visual automatic group affect analysis. *IEEE Transactions on Affective Computing*, 14(2):1056–1069, 2021.

[15] S. Ottl, S. Amiriparian, M. Gerczuk, et al. Group-level speech emotion recognition utilising deep spectrum features. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 821–826, 2020.

[16] X. Huang and J. Xu. Multimodal spatiotemporal semisupervised transformer network for video-based group-level emotion recognition. *IEEE Transactions on Computational Social Systems*, 2025.

[17] K. G. Quach, N. Le, C. N. Duong, I. Jalata, K. Roy, and K. Luu. Non-volume preserving-based fusion to group-level emotion recognition on crowd videos. *Pattern Recognition*, 128:108646, 2022. doi: 10.1016/j.patcog.2022.108646.

Journal Pre-proof

Journal Pre-proof

[18] M. Sun, J. Li, H. Feng, et al. Multi-modal fusion using spatio-temporal and static features for group emotion recognition. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 835–840, 2020.

[19] X. Ma, X. Huang, and H. Li. Multi-view driven and modality-adaptive model for emotion recognition in conversation. *Knowledge-Based Systems*, page 114839, 2025.

[20] Y. Wang, J. Wu, P. Heracleous, et al. Implicit knowledge injectable cross attention audiovisual model for group emotion recognition. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 827–834, 2020.

[21] L. Evtodienko. Multimodal end-to-end group emotion recognition using cross-modal attention. *arXiv preprint arXiv:2111.05890*, 2021.

[22] Y. H. H. Tsai, S. Bai, P. P. Liang, et al. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, 2019.

[23] A. Malhotra, G. Sharma, R. Kumar, et al. Social event context and affect prediction in group videos. In *2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 1–8. IEEE, 2023.

[24] W. Han, H. Chen, A. Gelbukh, et al. Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pages 6–15, 2021.

[25] A. Radford, J. W. Kim, C. Hallacy, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[26] B. Elizalde, S. Deshmukh, M. Al Ismail, et al. CLAP: learning audio concepts from natural language supervision. In *ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[27] Z. Lian, L. Sun, H. Sun, et al. GPT-4V with emotion: A zero-shot benchmark for generalized emotion recognition. *Information Fusion*, 108:102367, 2024.

[28] W. Hong, W. Yu, X. Gu, et al. GLM-4.1 V-Thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

[29] B. Yu, J. Wang, C. Li, et al. Graph-prototype distillation with prototype-guided contrastive training for multimodal emotion recognition in conversations. *Knowledge-Based Systems*, page 115484, 2026.

[30] Y. Shou, T. Meng, W. Ai, et al. Multimodal large language models meet multimodal emotion recognition and reasoning: A survey. *arXiv preprint arXiv:2509.24322*, 2025.

[31] Z. Li, B. Xu, C. Zhu, et al. CLMLF: A contrastive learning and multi-layer fusion method for multimodal sentiment detection. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2282–2294, 2022.

[32] K. Slogrove and D. Van Der Haar. Group emotion recognition in the wild using pose estimation and LSTM neural networks. In *2022 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD)*, pages 1–6. IEEE, 2022. doi: 10.1109/icABCD54961.2022.9856227.

[33] A. Augusma, D. Vaufreydaz, and F. Letué. Multimodal group emotion recognition in-the-wild using privacy-compliant features. In *Proceedings of the 25th International Conference on Multimodal Interaction*, pages 750–754, 2023. doi: 10.1145/3577190.3616546.

[34] P. Ekman. An argument for basic emotions. *Cognition & Emotion*, 6 (3-4):169–200, 1992.

[35] H. Tajfel, J. Turner, W. G. Austin, et al. An integrative theory of intergroup conflict. In *Intergroup Relations: Essential Readings*, pages 94–109. 2001.

[36] P. Ekman. Universals and cultural differences in facial expressions of emotion. In *Nebraska Symposium on Motivation*. University of Nebraska Press, 1971.

[37] L. F. Barrett, B. Mesquita, and M. Gendron. Context in emotion perception. *Current Directions in Psychological Science*, 20(5):286–290, 2011.

[38] R. Banse and K. R. Scherer. Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, 70(3):614, 1996.

[39] J. A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161, 1980.

[40] K. R. Scherer. The dynamic architecture of emotion: Evidence for the component process model. *Cognition and Emotion*, 23(7):1307–1351, 2009.

[41] R. Collins. *Interaction Ritual Chains*. Princeton University Press, 2014.

[42] J. R. Pinto, T. Gonçalves, C. Pinto, et al. Audiovisual classification of group emotion valence using activity recognition networks. In *2020 IEEE 4th International Conference on Image Processing, Applications and Systems (IPAS)*, pages 114–119. IEEE, 2020.

[43] A. Petrova, D. Vaufreydaz, and P. Dessus. Group-level emotion recognition using a unimodal privacy-safe non-individual approach. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 813–820, 2020.

[44] A. V. Savchenko, L. V. Savchenko, and I. Makarov. Classifying emotions and engagement in online learning based on a single facial expression recognition neural network. *IEEE Transactions on Affective Computing*, 13(4):2132–2143, 2022.

Journal Pre-proof