frontiers in Artificial Intelligence

TYPE Original Research
PUBLISHED 17 November 2026
DOI 10.3389/frai.2026.1984257

Check for updates

OPEN ACCESS

EDITED BY
Simone Bonechi,
University of Siena, Italy

REVIEWED BY
Junhwa Kim,
Konyang University - Daejeon Medical Campus, Republic of Korea
Kaiwen Zheng,
University of Glasgow, United Kingdom

*CORRESPONDENCE
Livia Del Gaudio
livia.delgaudio@unimore.it

RECEIVED 31 August 2026
REVISED 14 October 2026
ACCEPTED 29 October 2026
PUBLISHED 17 November 2026

CITATION
Del Gaudio L, Cuculo V and Cucchiara R (2026) From action units to grounded explanations: Multiple Instance Learning and Vision-Language Reasoning for video FER. *Front. Artif. Intell.* 9:1984257. doi: 10.3389/frai.2026.1984257

COPYRIGHT
© 2026 Del Gaudio, Cuculo and Cucchiara. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

From action units to grounded explanations: Multiple Instance Learning and Vision-Language Reasoning for video FER

Livia Del Gaudio*, Vittorio Cuculo and Rita Cucchiara

Department of Engineering "Enzo Ferrari", University of Modena and Reggio Emilia, Modena, Italy

Facial expressions alone are often insufficient to infer emotion: human observers routinely use scene, body and social context to disambiguate facial cues. Recent multimodal frameworks built on vision-language and large language models (VLMs/LLMs) reason jointly over facial and contextual evidence, but they typically process entire videos. This is computationally costly, and the few emotionally expressive frames are diluted among many uninformative ones. We introduce Reasoning from Facial Evidence for Emotion Recognition (ReFER), which focuses multimodal reasoning on the most relevant moments. A Multiple Instance Learning (MIL) module predicts the emotion from an Action Unit (AU) sequence and scores each frame's relevance. A VLM then describes the visual context of the top-scoring frames. Finally, an instruction-tuned LLM fuses the MIL prediction, the AU measurements and the context description into a final label and a natural-language explanation. Beyond accuracy, we analyze which inputs actually drive the fused prediction. Across four in-the-wild video facial expression recognition (FER) benchmarks and five open-weight fusion LLMs, the best fusion models outperform the AU-only MIL baseline on seven of eight dataset-metric pairs, with the largest gains on AFEW and DFEW. Component ablations show that the initial MIL prediction, rather than the contextual description, is the dominant input to the fusion stage. A label-corruption analysis further reveals that all fusion models anchor strongly on the supplied initial prediction, adopting corrupted labels in 82.6%–99.9% of cases. These results show that an explanation consistent with the available evidence does not imply that the evidence faithfully determines the prediction and points to the need for fusion mechanisms explicitly designed to reconcile conflicting multimodal evidence.

KEYWORDS
affective computing, face analysis, multimodal emotion recognition, Multiple Instance Learning, Vision-Language Models

1 Introduction

The interpretation of human emotions is inherently contextual. Although facial expressions constitute one of the most informative channels for affective communication, psychological evidence has consistently demonstrated that observers rarely infer emotions from facial configurations alone. Instead, emotional meaning emerges from the interaction between facial signals, body posture, environmental cues, social interactions and situational knowledge (Carroll and Russell, 1996; Aviezer et al., 2008, 2012; Barrett et al., 2011; Wieser and Brosch, 2014; Boccignone et al., 2020). The same facial expression may

therefore be perceived as conveying different emotions depending on the surrounding scene and the observer's interpretation of the ongoing situation. Such findings challenge the long-standing assumption that emotion recognition can be reliably achieved through facial analysis alone.

This limitation becomes particularly evident in real-world affective computing scenarios. In unconstrained environments, facial observations are often ambiguous, partially occluded, or insufficient to explain the emotional state of an individual. For instance, widened eyes and raised eyebrows may correspond to fear, surprise, excitement, or anticipation, depending on contextual circumstances. Similarly, facial muscle activations associated with smiling can reflect joy, embarrassment, politeness, sarcasm, or even frustration. Consequently, systems that rely exclusively on facial appearance may capture the physiological manifestation of emotion while failing to identify its underlying semantic meaning.

To address these limitations, recent research has increasingly incorporated contextual information into emotion recognition pipelines. Context-aware approaches demonstrated that scene understanding significantly improves affect recognition by exploiting information related to surrounding objects, environmental characteristics, and interpersonal interactions (Kosti et al., 2017; Colaco and Han, 2025). These findings align with contemporary psychological theories suggesting that emotional perception operates as an inferential process in which facial cues are interpreted through the lens of contextual appraisal (Barrett et al., 2011; Wieser and Brosch, 2014). As a result, the central challenge is no longer merely identifying facial patterns associated with emotions, but understanding how facial and contextual evidence interact to support an affective judgment.

At the same time, the emergence of Vision-Language Models (VLMs) and Large Language Models (LLMs) has profoundly transformed multimodal affective computing. Rather than directly mapping observations to emotion labels, these models enable explicit reasoning over multimodal evidence and the generation of human-interpretable explanations. Recent frameworks such as *Explainable Multimodal Emotion Reasoning* (EMER) introduced the concept of affective reasoning as a task that combines prediction and explanation (Lian et al., 2023). Building upon this perspective, Zhang H. et al. (2026) proposed cue-level emotional perception and reasoning mechanisms capable of identifying fine-grained emotional evidence before producing affective interpretations. Similarly, Wu et al. (2026) reformulated facial emotion analysis as a reasoning problem by jointly modeling Action Units (AUs), emotion recognition, and AU-based explanations, demonstrating that aligning reasoning and recognition improves both robustness and interpretability. More recently, emotion-aware reasoning frameworks (Rha et al., 2026) emphasized the necessity of maintaining consistency between generated explanations and predicted emotional states, while evidence-centric approaches (Zhao et al., 2026) showed that separating evidence extraction from reasoning substantially improves multimodal emotion understanding under ambiguous or conflicting conditions. These studies collectively indicate a transition from discriminative classification toward explainable affective reasoning systems capable of explicitly linking observations to decisions.

Despite these advances, two important challenges remain unresolved. First, most multimodal reasoning architectures process entire video sequences, resulting in significant computational costs and the inclusion of large amounts of emotionally irrelevant information. Second, contextual reasoning is frequently performed independently from the mechanisms responsible for identifying emotionally salient facial evidence, leading to suboptimal integration between facial dynamics and scene-level semantics.

Our previous works (Del Gaudio et al., 2026a,b) addressed the first part of this problem through a Multiple Instance Learning (MIL) framework operating on Facial Action Units. By representing a video as a collection of temporal AU instances, the model learns to identify the segments that contribute most strongly to the final emotional prediction while preserving temporal interpretability. The resulting attention mechanism provides an implicit localization of emotionally informative moments without requiring frame-level annotations. Such a property offers a unique opportunity for multimodal reasoning: rather than analyzing every frame of a video, computationally expensive foundation models can focus exclusively on the temporal regions that have already been identified as affectively relevant.

Motivated by this observation, we introduce Reasoning from Facial Evidence for Emotion Recognition (ReFER), a framework that combines AU-based temporal modeling, localized contextual descriptions and language-model-based evidence fusion. First, the MIL module assigns temporal relevance scores to the AU sequence, from which a small number of candidate observations are selected. A VLM then describes the visual context surrounding the selected observation, including the scene, salient objects and social interactions. Finally, an instruction-tuned LLM combines the initial MIL prediction, the corresponding AU measurements and the contextual description to produce a final emotion prediction and a natural-language explanation. This provides a computationally efficient interface between temporal AU modeling and foundation models, while retaining explicit access to the facial and contextual evidence supplied to the fusion stage. However, whether the selected observations are more informative than arbitrary frames, and whether the fusion LLM genuinely uses the supplied evidence rather than following the initial prediction, remain empirical questions. We therefore evaluate not only recognition performance and explanation quality, but also the contribution of each evidence source, the effectiveness of attention-guided frame selection, and the susceptibility of the fusion models to label anchoring.

The main contributions of this work are three-fold:

- We introduce ReFER, a framework that extends AU-based temporal emotion recognition with localized visual context and structured LLM-based fusion, producing both emotion predictions and natural-language explanations.
- We evaluate the framework across four in-the-wild video FER benchmarks and five open-weight fusion LLMs, analyzing both recognition performance and explanation quality through deterministic AU-grounding metrics and independent LLM judges.
- We investigate how the fusion models use the supplied evidence through component ablations, attention-guided

vs. random frame selection and a label-corruption intervention.

2 Related work

2.1 Contextual emotion perception

The importance of contextual information in emotion perception has been extensively investigated in psychology and cognitive science. Contrary to early views that emotional meaning could be directly inferred from facial configurations, a substantial body of evidence suggests that observers integrate facial cues with contextual information when interpreting affective states.

A seminal contribution by Carroll and Russell (1996) showed that identical facial expressions can be interpreted as conveying different emotions, depending on the accompanying situational context. Their findings challenged the notion of a fixed mapping between facial configurations and emotional categories, suggesting instead that emotional judgments emerge from an interaction between perceptual cues and contextual knowledge.

Subsequent studies further highlighted the dominant role of contextual information under ambiguous conditions. Aviezer et al. (2008) demonstrated that contextual body and scene information can substantially alter the perceived meaning of facial expressions. Indeed, in a later study, the same authors showed that body cues often outweigh facial information when observers interpret highly intense emotional states (Aviezer et al., 2012). More recently, psychological constructionist theories have proposed that emotion perception is fundamentally a context-sensitive inferential process. Barrett et al. (2011) argued that observers actively construct emotional interpretations by combining sensory information with contextual knowledge and prior experience. Similarly, Wieser and Brosch (2014) reviewed evidence demonstrating that contextual influences affect multiple stages of affective face processing, ranging from visual attention to higher-level emotional categorization.

Taken together, these studies suggest that human emotion perception relies on a hierarchical integration process in which facial signals provide affective evidence while contextual information supplies semantic interpretation. This perspective has become increasingly influential in affective computing, motivating computational systems that move beyond isolated facial analysis toward context-aware and reasoning-oriented emotion understanding.

2.2 Context-aware emotion recognition

Motivated by the psychological evidence discussed above, researchers have increasingly explored context-aware approaches for automatic emotion recognition. One of the most influential contributions in this direction is EMOTIC (Kosti et al., 2017), which introduced large-scale emotion recognition in context and demonstrated that scene information provides complementary affective cues beyond facial appearance and body posture. Subsequent approaches focused on improving the representation of contextual information. CAER-Net proposed a two-stream architecture that separately encodes facial and contextual features and fuses them through adaptive attention mechanisms (Lee et al., 2019). Building upon this idea, EmotiCon introduced a multimodal framework inspired by Frege's Context Principle, integrating facial cues, scene semantics, social interactions, and depth information to model multiple interpretations of context simultaneously (Mittal et al., 2020).

Recent studies have explored more advanced representations of contextual relationships. Yang et al. (2022) proposed a multiple-context framework combining facial appearance, scene information, interpersonal relationships, and agent-object interactions. Hierarchical scene-graph approaches further incorporated object-level and semantic relationships into emotion prediction, showing that relational reasoning improves performance over global scene descriptors alone (Wu et al., 2022). Similarly, visual relationship detection methods explicitly modeled interactions between individuals and surrounding objects, demonstrating that contextual relevance depends not only on the presence of objects but also on their semantic relationship with the target subject (Hoang et al., 2021). More recently, transformer-based approaches such as CD-Net leveraged long-range dependencies among facial, bodily and contextual cues to improve robustness in complex scenes (Wang Z. et al., 2022), while TDRCer introduced interaction graphs combining gaze, depth and object relationships to model contextual emotional dynamics (Zhang et al., 2024).

Despite their success, most context-aware emotion recognition systems remain fundamentally discriminative. Contextual information is generally fused with visual features prior to classification, while little attention is devoted to explaining how contextual evidence contributes to the final emotional interpretation. Furthermore, contextual analysis is typically performed over entire images or videos, potentially introducing irrelevant information and unnecessary computational complexity.

2.3 Multimodal emotion reasoning with foundation models

The emergence of Vision-Language Models (VLMs) and Large Language Models (LLMs) has recently transformed affective computing from a recognition-oriented paradigm into a reasoning-oriented one. Instead of directly predicting emotion labels, these models seek to identify emotionally relevant evidence, establish relationships among multimodal cues, and generate explicit explanations supporting their predictions. Among the first works to formalize this perspective, Explainable Multimodal Emotion Reasoning (EMER) introduced a benchmark and framework in which emotion recognition is coupled with explanation generation, encouraging models to justify their predictions through multimodal evidence (Lian et al., 2023). More recent approaches have further emphasized the importance of explicit reasoning processes. XEmoGPT (Zhang H. et al., 2026) proposed cue-level emotional perception and reasoning mechanisms that first identify fine-grained emotional evidence and subsequently generate affective interpretations grounded

in the observed cues. Facial-R1 (Wu et al., 2026) extended this paradigm to facial analysis by jointly modeling Action Units, emotion recognition, and reasoning, showing that alignment between evidence and explanation improves robustness and interpretability.

Recent research has further emphasized the importance of consistency between explanations and predictions. Rha et al. (2026) introduced explicit verification mechanisms that assess whether generated rationales actually support the predicted emotional state, while R1-Omni (Zhao et al., 2025) leveraged reinforcement learning to align multimodal emotional reasoning with human preferences and reduce hallucinated evidence. In the video domain, VidEmo (Zhang Z. et al., 2026) introduced a family of video emotion foundation models trained through curriculum emotion learning followed by affective-tree reinforcement learning, unifying attribute perception, expression analysis, and high-level emotional understanding in a stage-wise reasoning process, and further contributed Emo-CFG, a large-scale instruction dataset with explainable rationales for emotion understanding. A complementary research direction focuses on evidence-centric multimodal reasoning. Frameworks such as SABER-LLM (Zhao et al., 2026) proposed a "perceive-then-reason" paradigm in which multimodal evidence is first extracted and structured before being processed by reasoning modules. This separation between perception and reasoning was shown to improve robustness under emotionally ambiguous conditions while enhancing interpretability. Other recent works have also begun to address the computational cost of multimodal emotion reasoning. Nano-EmoX (Huang et al., 2026) proposes a compact multimodal language model for emotion perception, understanding and interaction, while Light-MER (Zheng et al., 2026) transfers multimodal affective capabilities from a larger teacher to a sub-billion-parameter student through distillation followed by multi-reward reinforcement refinement. These works highlight a complementary trend toward more computationally practical foundation-model-based emotion understanding. However, they reduce cost at the model level, by shrinking or distilling the reasoning model, rather than at the input level. Indeed, despite their promising results, existing reasoning-based frameworks, including these efficient variants, typically process complete multimodal inputs and lack mechanisms for identifying emotionally informative temporal regions before reasoning. Consequently, foundation models are often required to analyze large amounts of redundant information, resulting in high computational costs and potentially weaker evidence grounding.

Our work aims to bridge the gap between these research directions. Unlike traditional context-aware methods, contextual information is not indiscriminately fused with facial features but is selectively analyzed only within temporally localized affective segments. Unlike AU-based approaches, contextual evidence is explicitly incorporated into the decision-making process. Finally, unlike recent reasoning-based foundation models, the proposed framework exploits a Multiple Instance Learning mechanism to guide contextual perception, enabling efficient Vision-Language Model usage and grounding Large Language Model reasoning in temporally localized facial evidence.

3 Proposed method

3.1 Framework overview

Emotion recognition in unconstrained videos requires reasoning over multiple sources of evidence. While facial expressions provide a direct manifestation of affective states, psychological and computational studies have shown that emotion perception is strongly influenced by contextual information, including body posture, interpersonal interactions, surrounding objects and scene characteristics (Bhattacharya et al., 2021; Mittal et al., 2021). Consequently, relying exclusively on facial information may overlook cues that are relevant for disambiguating visually similar expressions. At the same time, not all portions of a video contribute equally to emotion recognition. Emotional evidence is often concentrated in a limited number of salient temporal segments, suggesting that effective recognition systems should identify and focus on the most informative observations rather than processing all frames equally.

Motivated by these considerations, ReFER is built around three principles. First, facial behavior is represented using the Facial Action Coding System (FACS) descriptors, providing an interpretable characterization of expression dynamics in terms of observable facial muscle activations. Second, a temporal Multiple Instance Learning (MIL) formulation is employed to identify the portions of the video that contain the most discriminative emotional evidence. This allows the framework to focus subsequent analysis on a small subset of salient observations. Third, contextual cues associated with the selected observations are integrated with facial evidence through a reasoning stage that produces both a final emotion prediction and a textual explanation grounded in the available evidence.

The resulting framework separates evidence acquisition from evidence integration, enabling predictions to remain interpretable while incorporating complementary contextual information. A schematic overview of the framework is shown in Figure 1.

3.2 Facial dynamics modeling

3.2.1 Action Unit representation

Given a video $\mathcal{V} = \{f_t\}_{t=1}^{T}$, each frame is represented through a vector of facial Action Unit (AU) intensities $\mathbf{a}_t \in [0,1]$, yielding the multivariate temporal sequence $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_T] \in \mathbb{R}^{T \times A}$. Action Units provide an interpretable description of facial activity by decomposing facial expressions into elementary muscular actions. Unlike appearance-based representations, this formulation allows predictions to be directly linked to semantically meaningful facial behaviors.

3.2.2 Multiple Instance Learning

Video emotion recognition naturally fits the Multiple Instance Learning (MIL) paradigm. As introduced in our previous work (Del Gaudio et al., 2026a), each video is treated as a bag of temporally ordered observations while supervision is available

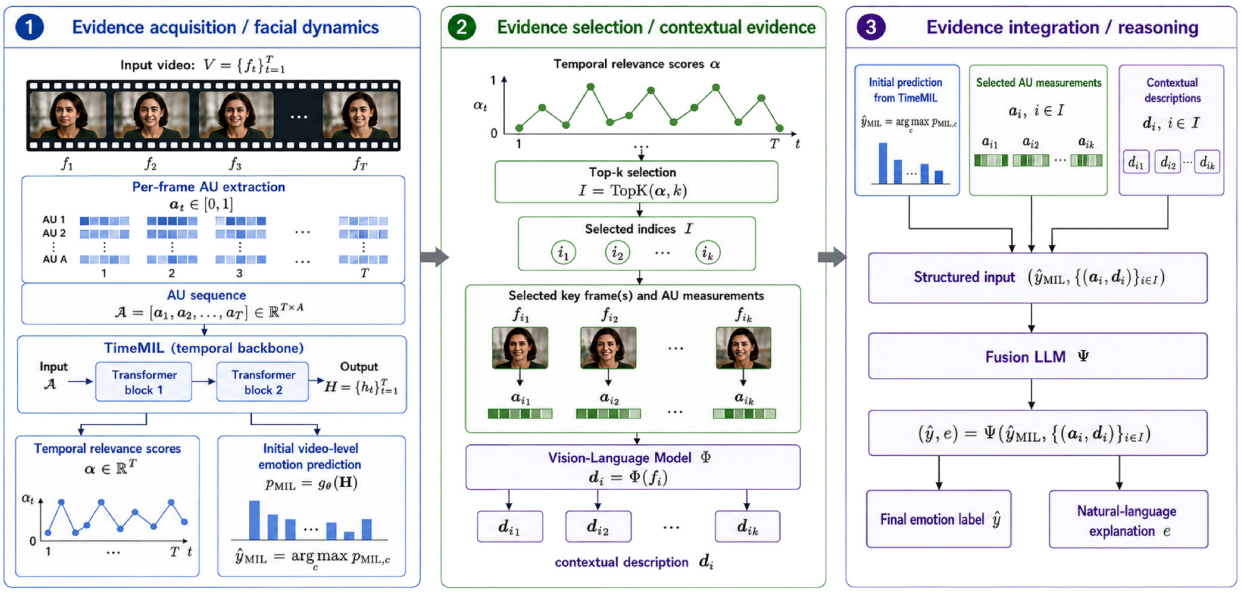

FIGURE 1
ReFER pipeline overview. Given an input video, Action Units (AUs) are extracted for each frame. The AU sequence is processed by a Multiple Instance Learning (MIL) module that produces a video-level emotion prediction and temporal relevance scores (1). The frames with the highest relevance are selected and described by a Vision-Language Model (VLM) to obtain a contextual description (2). Finally, an instruction-tuned LLM fuses the initial prediction, AU measurements, and the contextual description to output the final emotion label and a natural language explanation (3). Images were generated using OpenAI ChatGPT Images 2.0.

only at the video level. Under this formulation, discriminative evidence is not assumed to be uniformly distributed throughout the sequence; instead, only a subset of observations may be strongly associated with the target emotion.

The AU sequence is therefore encoded into a set of latent temporal representations $\mathbf{H} = \{\mathbf{h}_t\}_{t=1}^{T}$, which are used to estimate a video-level emotion distribution according to Equation 1:

$$\mathbf{p}_{\mathrm{MIL}} = g_\theta(\mathbf{H}), \qquad \hat{y}_{\mathrm{MIL}} = \arg\max_c \mathbf{p}_{\mathrm{MIL},c}. \tag{1}$$

In addition to the video-level prediction, the MIL formulation provides a measure of temporal relevance that can be used to identify the observations that contributed most strongly to the decision.

3.2.3 Salient temporal evidence selection

The temporal relevance scores produced by the MIL stage are used to localize emotionally informative segments within the video.

Let $\boldsymbol{\alpha} \in \mathbb{R}^T$ denote the temporal relevance scores associated with the sequence. The indices corresponding to the top-k observations are selected as defined in Equation 2:

$$\mathcal{I} = \mathrm{TopK}\left(\boldsymbol{\alpha}, k\right), \tag{2}$$

This procedure can be interpreted as an evidence-localization mechanism: rather than treating all frames as equally informative, the framework focuses on a small set of temporal locations that are expected to contain the strongest emotional evidence. Restricting subsequent processing to these observations also improves computational efficiency while preserving interpretability.

3.3 Contextual evidence extraction

As introduced in the previous sections, facial behavior alone may not fully characterize the emotional meaning of an event. Research on context-aware emotion recognition has shown that body language, social interactions, objects, and environmental characteristics can substantially influence emotion perception. To capture this complementary information, contextual evidence is extracted from the observations identified by the temporal selection stage.

For each selected index $i \in \mathcal{I}$, a contextual description $\mathbf{d}_i = \Phi(f_i)$ is obtained from the corresponding frame, where $\Phi(\cdot)$ denotes a contextual perception function that summarizes relevant visual information surrounding the facial expression.

Each selected observation is therefore represented by two complementary sources of evidence: (i) facial evidence, represented by the Action Unit measurements $\mathbf{a}_i$; (ii) contextual evidence, represented by the description $\mathbf{d}_i$. These paired observations are subsequently provided to the fusion stage together with the initial emotion prediction.

3.4 Evidence fusion and explanation generation

The final stage integrates facial and contextual evidence to produce both an emotion prediction and a human-readable explanation.

Given the selected observations, the framework constructs a structured representation containing $\{(\mathbf{a}_i, \mathbf{d}_i)\}_{i \in \mathcal{I}}$, together with the initial video-level prediction $\hat{y}_{\mathrm{MIL}}$.

TABLE 1 Overview of the evaluation benchmarks.

Dataset	#Videos	#Classes	Split protocol	Source domain
AFEW	1,809	7	Official train/val split	Movies/TV
DFEW	16,372	7	Five-fold cross-validation	Movies, in-the-wild clips
MAFW	10,045	11	Five-fold cross-validation	Movies/TV, multimodal
FERV39k	38,935	7	Official train/val/test split	Web videos, four scenarios

#Classes reports the number of categorical emotion labels used at evaluation time (MAFW uses an extended 11-class taxonomy; the remaining datasets use the standard seven basic emotions).

The final prediction is obtained through an evidence-fusion function, as expressed in Equation 3:

$$(\hat{y}, e) = \Psi\left(\hat{y}_{\text{MIL}}, \{(\mathbf{a}_i, \mathbf{d}_i)\}_{i\in\mathcal{I}}\right), \tag{3}$$

where $\hat{y}$ denotes the final emotion label and e denotes the associated explanation.

The role of the fusion stage is not merely to aggregate information, but to reason jointly over multiple evidence sources. The initial prediction provides a global hypothesis derived from the temporal facial dynamics, while the selected facial and contextual observations provide localized supporting evidence. The generated explanation is required to justify the final decision by explicitly referring to the available observations, thereby improving transparency and interpretability.

4 Experiments

4.1 Datasets and evaluation protocol

We evaluate ReFER on four in-the-wild video-based facial expression recognition (FER) benchmarks (see Table 1): AFEW (Dhall et al., 2014), DFEW (Jiang et al., 2020), FERV39k (Wang Y. et al., 2022) and MAFW (Liu et al., 2022). These datasets were selected because they span complementary sources of variability relevant to in-the-wild affective computing: movie clips (AFEW, MAFW), short user-generated and web videos (DFEW) and large-scale crowd-sourced footage (FERV39k), and because they are the standard benchmarks against which FER methods, including our prior pipeline (Del Gaudio et al., 2026a), are conventionally compared.

AFEW and FERV39k are evaluated on their official splits. For DFEW we follow the standard protocol and report the mean and standard deviation over the five official cross-validation folds. For MAFW we report results on the first official fold.

4.2 Implementation details

Following our previous work (Del Gaudio et al., 2026a), facial activity is represented using a fixed set of 20 Facial Action Units (AUs) extracted with Py-Feat (Cheong et al., 2023). AU extraction is performed once as an offline preprocessing step, before model training, and the resulting per-frame AU intensity sequences are cached and reused throughout all subsequent training and inference stages. The cached AU sequences are z-normalized per video and provided to TimeMIL (Chen et al., 2024), which serves as the temporal Multiple Instance Learning (MIL) backbone. TimeMIL is trained independently on each dataset using only video-level emotion supervision, following a bag-of-instances MIL formulation in which each video is treated as a bag of per-frame AU instances and only the video-level label is observed in training.

Once trained, TimeMIL is frozen and used to produce (i) the initial video-level emotion prediction $\hat{y}_{\text{MIL}}$ and (ii) the temporal relevance (attention) scores over the AU sequence, which are used for salient-frame selection. Unless otherwise stated, the number of selected temporal observations is set to $k = 1$, corresponding to the highest-ranked temporal location according to the attention distribution; the effect of this choice relative to random frame selection and for $k > 1$ is examined in Section 5.3.

The video frames corresponding to the selected temporal locations are processed by InternVL3 (Zhu et al., 2025) to generate a free-text contextual description of the visual scene (e.g., setting, social context, salient objects and interactions), which is paired with the AU measurements extracted at the same temporal locations and supplied to the fusion LLM. The context-extraction stage (InternVL3, frame selection and prompt template) is kept fixed across all experiments so that variation in downstream results can be attributed exclusively to the evidence-fusion stage.

All VLM and LLM components are used in inference mode with frozen, publicly released pretrained weights; no additional fine-tuning is performed on any dataset or fusion model. The same context-extraction and evidence-fusion prompt templates are used across all four datasets and all five fusion LLMs, with one exception: for MAFW, the list of candidate emotion categories provided to the fusion LLM is adapted to its extended 11-class label space. The fusion LLM receives three sources of evidence: (i) the initial prediction $\hat{y}_{\text{MIL}}$ from TimeMIL, (ii) the AU measurements at the selected temporal location, and (iii) the InternVL3-generated contextual description; it is prompted to return a final emotion label together with a concise natural-language explanation grounding that label in the supplied evidence.

For the sake of reproducibility, the context-extraction and evidence-fusion prompt templates used are reported in the Supplementary material.

4.3 Fusion models

The evidence-fusion stage is instantiated with several instruction-tuned large language models (LLMs) in order to assess the robustness of ReFER with respect to the reasoning component, rather than to a single choice of LLM.

We evaluate three general-purpose open-weight models of comparable scale: Qwen2.5-7B-Instruct (Yang et al., 2024), Mistral-7B-Instruct (Jiang et al., 2023), and Llama-3.1-8B-Instruct (Grattafiori et al., 2024). In addition, we consider EmoLlama-chat-7B (Liu et al., 2024), a model instruction-tuned specifically for emotion understanding and its corresponding non-affective base model, Llama-2-7B-chat (Touvron et al., 2023). Contrasting EmoLlama-chat-7B with Llama-2-7B-chat provides a controlled comparison in which architecture, tokenizer and parameter count are matched while the fine-tuning objective differs, isolating the contribution of affect-specific instruction tuning from purely architectural factors.

4.4 Evaluation protocol

We evaluate ReFER along two complementary dimensions: emotion recognition performance and the quality of the generated explanations.

4.4.1 Recognition performance

Recognition performance is measured using Weighted Average Recall (WAR) and Unweighted Average Recall (UAR), consistent with standard evaluation protocols for in-the-wild video FER. Denoting by C the number of emotion classes and by Recall_c the per-class recall, WAR corresponds to overall (sample-weighted) accuracy and is computed using Equation 4:

$$\text{WAR} = \frac{\sum_{c=1}^{C} n_c \cdot \text{Recall}_c}{\sum_{c=1}^{C} n_c}, \tag{4}$$

where n_c is the number of test samples of class c, while UAR is the unweighted mean of per-class recalls, as defined in Equation 5:

$$\text{UAR} = \frac{1}{C} \sum_{c=1}^{C} \text{Recall}_c. \tag{5}$$

UAR is reported alongside WAR because all four benchmarks are class-imbalanced; UAR is therefore the more informative metric for assessing performance on minority emotion categories.

4.4.2 AU grounding

We check whether the AUs mentioned in each explanation are supported by the AU measurements given to the fusion model and by the Facial Action Coding System (FACS) (Ekman and Friesen, 1978). All checks are deterministic. We call every mention of an AU in an explanation a *citation*. A citation may give only the AU code (e.g., "AU12") or the code together with its name (e.g., "AU12, lip corner puller"). We report four complementary measures.

FACS consistency asks whether the cited AUs are the ones FACS associates with the predicted emotion. It is the fraction of citations whose AU belongs to the canonical AU pattern of that emotion (e.g., AU6 and AU12 for happiness). To keep the protocol identical across datasets, we compute it only for the seven emotions shared by all four benchmarks (anger, disgust, fear, happiness, neutral, sadness, and surprise). MAFW predictions of contempt, anxiety, helplessness, and disappointment are therefore excluded from this measure, but not from the other three.

Naming accuracy asks whether the AU names given in the explanation are correct. It is the fraction of named citations whose name matches the FACS definition of the cited code. Citations that give only a code are not counted.

Cited-AU percentile asks whether the explanation cites AUs that are actually strongly active. At the selected frame, we rank each cited AU's intensity among the 20 tracked AUs and convert the rank to a percentile, where 100% means it is the most intense AU in that frame. If several AUs are cited, their percentiles are averaged. Higher values mean the explanation focuses on the most prominent facial movements. Note that this compares AUs to one another within a single frame, not over time.

Finally, *Top-1 AU citation* asks whether the explanation mentions the single most active AU. It is the fraction of explanations that cite the AU with the highest intensity at the selected frame. Explanations that cite no AU count as failures.

4.4.3 Semantic properties

Semantic properties that cannot be assessed deterministically are evaluated using two independent LLM judges, Qwen2.5-32B-Instruct and Llama-3.3-70B-Instruct.

We note that one judge (Qwen2.5-32B-Instruct) shares a model family with one of the evaluated fusion LLMs (Qwen2.5-7B-Instruct); potential self-preference bias arising from this overlap is examined explicitly below. We assess (i) *context consistency*, i.e., whether the generated explanation is supported by the VLM-derived contextual evidence, and (ii) *prediction coherence*, i.e., whether the explanation is logically consistent with the final emotion prediction. Each dimension is scored independently on a five-point Likert-type scale using predefined textual anchors (1 = contradicts the evidence, 5 = fully supported).

The complete system and user prompt templates, including the scoring anchors and required output format are provided in Supplementary material.

5 Results

Table 2 reports recognition performance for the MIL baseline and for the five Full-pipeline configurations across all four benchmarks. Averaged across fusion LLMs, the full pipeline improves over the MIL baseline on AFEW and DFEW for four of five fusion models, with Qwen2.5-7B yielding the largest gains (+1.41 WAR/+1.35 UAR on AFEW; +1.06 WAR/+1.17 UAR on DFEW). Gains on MAFW and FERV39k are smaller and less consistent across fusion models, suggesting that the benefit of contextual evidence and LLM-based fusion is not uniform across datasets, plausibly reflecting differences in how strongly scene context disambiguates the emotion categories in each corpus (e.g., MAFW's finer-grained 11-class taxonomy may require context that is not well-captured by a single salient frame). A paired

TABLE 2 Main results across four in-the-wild video FER benchmarks.

			AFEW		DFEW		MAFW		FERV39k	
Method	Fusion LLM	Modality	WAR	UAR	WAR	UAR	WAR	UAR	WAR	UAR
MIL (baseline)	–	AU	38.90	36.46	52.43 ± 1.28	42.17 ± 1.25	32.25	22.45	37.91	26.80
Full (proposed)	Qwen2.5-7B	AU + context	**40.31**	**37.81**	**53.49 ± 1.38**	**43.34 ± 1.34**	33.18	22.72	37.86	**27.50**
Full (proposed)	Mistral-7B	AU + context	39.53	37.49	47.65 ± 1.27	40.78 ± 1.49	32.67	22.62	36.52	26.76
Full (proposed)	Llama-3.1-8B	AU + context	34.03	32.58	45.65 ± 1.20	39.75 ± 1.23	33.12	22.43	36.75	27.02
Full (proposed)	Llama-2-7B	AU + context	39.01	36.63	52.42 ± 1.28	42.16 ± 1.25	32.21	22.43	38.20	26.59
Full (proposed)	EmoLlama-7B	AU + context	39.27	36.86	52.80 ± 1.34	42.50 ± 1.27	**33.24**	**23.00**	**38.25**	26.78

MIL is the AU-based baseline without LLM fusion (Del Gaudio et al., 2026a). Full denotes the proposed pipeline, which combines the MIL prediction, AU measurements, and VLM-derived contextual descriptions through structured LLM fusion. AFEW and FERV39k are evaluated on their official splits, DFEW results are reported as mean ± standard deviation over the five official folds, and MAFW results refer to the first official fold. The best WAR and UAR per dataset among the Full-pipeline variants are shown in bold.

TABLE 3 AU grounding of the generated explanations, macro-averaged across AFEW, DFEW, MAFW, and FERV39k, with each dataset contributing equally.

Fusion LLM	FACS consistency (%)	Naming Acc. (%)	Cited AU percentile (%)	Top-1 cited AU (%)
Qwen2.5-7B	40.07	81.77	**78.48**	**73.78**
Mistral-7B	66.68	**94.50**	62.56	44.91
Llama-3.1-8B	**70.32**	93.18	68.42	47.68
Llama-2-7B	32.23	84.14	58.53	16.45
EmoLlama-7B	28.09	55.98	60.38	15.03

FACS consistency measures agreement between cited AUs and the canonical pattern of the predicted emotion and is evaluated on the seven emotion categories shared across all benchmarks. Naming Accuracy evaluates explicitly provided AU-name associations; Cited AU Percentile measures the relative intensity of cited AUs among the 20 AU channels at the selected frame; and Top-1 AU Cited reports whether the most active AU at that frame is mentioned. Dataset-specific results and evaluation denominators are reported in the Supplementary material. Bold values indicate the best results.

bootstrap analysis with 10,000 resamples (fold-stratified for DFEW) further confirms that the improvement from the MIL baseline to the Full ReFER configuration with Qwen2.5-7B is statistically reliable on DFEW for both WAR and UAR, and on at least one of the two metrics for each of the remaining datasets. No dataset shows a statistically reliable degradation on either metric. Additional results and confidence intervals are reported in the Supplementary material.

As a matter of fact, not all fusion LLMs benefit equally from the additional evidence: Llama-3.1-8B-Instruct underperforms the MIL baseline on AFEW (34.03 vs. 38.90 WAR) and DFEW (45.65 vs. 52.43 WAR), indicating that access to richer evidence does not guarantee improved recognition and that the fusion model's ability to weigh AU and contextual evidence appropriately is itself a bottleneck. In contrast, EmoLlama-chat-7B, fine-tuned specifically for emotion understanding, outperforms its non-affective base model, Llama-2-7B-chat, on three of four benchmarks (MAFW, FERV39k, and DFEW), providing initial evidence that affect-specific instruction tuning is beneficial for the fusion step even when the underlying architecture and evidence are held fixed.

Table 3 reports the deterministic AU-grounding metrics. Each metric is first computed independently for AFEW, DFEW, MAFW and FERV39k; dataset-level values are then macro-averaged so that each benchmark contributes equally. The results reveal complementary grounding behaviors across fusion models. Llama-3.1-8B achieves the highest FACS consistency (70.32%), while Mistral-7B achieves the highest naming accuracy (94.50%). Qwen2.5-7B, however, is more strongly grounded in the AU measurements available at the selected frame, achieving both the highest cited-AU intensity percentile (78.48%) and Top-1 AU citation rate (73.78%). This distinction indicates that consistency with canonical FACS patterns and grounding in sample-specific AU activations capture different properties of the generated explanations. Dataset-specific results and metric denominators are reported in the Supplementary material.

Table 4 reports the LLM-judged semantic evaluation on DFEW. Qwen2.5-7B receives the highest context-consistency and prediction-coherence scores from the Llama-3.3-70B judge (3.87 and 4.94, respectively) and the highest prediction-coherence score from the Qwen2.5-32B judge (4.61), while EmoLlama-7B receives the highest context-consistency score from the Qwen2.5-32B judge (3.45). We note that the Qwen2.5-32B judge assigns systematically lower absolute scores than the Llama-3.3-70B judge across all fusion models and both dimensions, which is consistent with either a stricter scoring tendency of that judge or a scale-calibration difference between the two judges rather than with a true difference in explanation quality; for this reason, we base cross-model comparisons primarily on the *relative ordering* of fusion models within each judge rather than on absolute score differences across judges. These semantic scores assess consistency with the VLM-derived contextual description and the final prediction supplied to the judge, rather than factual correctness with respect to the original video.

5.1 Computational cost

We additionally characterized the computational cost of the complete pipeline on a random subset of 100 DFEW videos. When

TABLE 4 Semantic evaluation of the explanations generated on DFEW using two independent LLM judges.

	Llama-3.3-70B judge		Qwen2.5-32B judge	
Fusion LLM	Context consistency	Prediction coherence	Context consistency	Prediction coherence
Qwen2.5-7B	**3.87 ± 0.75**	**4.94 ± 0.29**	3.41 ± 0.71	**4.61 ± 0.53**
Mistral-7B	3.77 ± 0.79	4.50 ± 0.57	3.40 ± 0.60	4.03 ± 0.34
Llama-3.1-8B	3.54 ± 0.89	4.73 ± 0.58	3.29 ± 0.68	4.39 ± 0.63
Llama-2-7B	3.32 ± 1.04	4.53 ± 0.64	3.32 ± 0.65	4.15 ± 0.63
EmoLlama-7B	3.59 ± 1.09	4.80 ± 0.56	**3.45 ± 0.70**	4.47 ± 0.62

Context consistency evaluates whether the explanation is supported by the VLM-derived contextual evidence, while prediction coherence measures its consistency with the final emotion prediction. Scores range from 1 to 5 and are reported as mean ± standard deviation. Bold values indicate the best results.

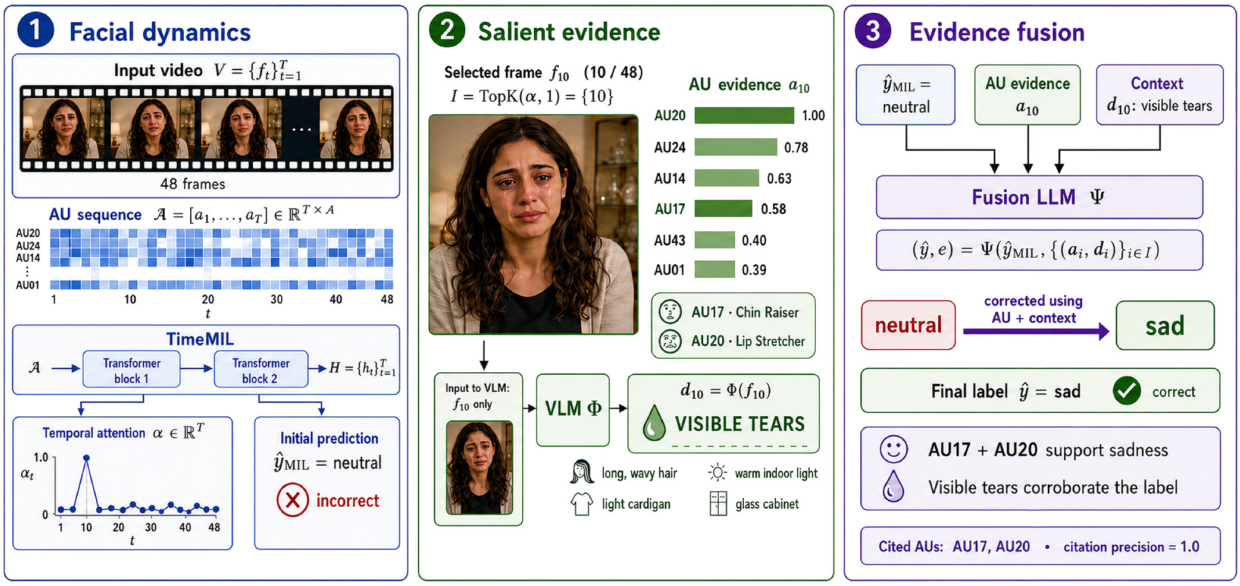

FIGURE 2
ReFER prediction-correction example for video ID 188 from DFEW. MIL model assigns the highest temporal relevance to frame 10/48 but initially predicts neutral. At the selected frame, AU17 and AU20 provide facial evidence supporting sadness, while VLM identifies visible tears in the contextual description. LLM integrates the initial prediction, AU measurements and contextual evidence, revising the final label to the correct class, sad. The displayed frames are generated visualizations and do not reproduce the original DFEW footage, as permission allows use of the dataset but not reproduction or distribution of its data. Images were generated using OpenAI ChatGPT Images 2.0.

starting from raw video, end-to-end processing required 26.02 ± 8.11 s/video, with the offline AU extraction stage accounting for most of the latency (21.91 ± 8.24 s, 84%). Once cached AU features are available, the online ReFER stages are substantially cheaper: TimeMIL inference required 0.028±0.011 s/video, while InternVL3 context generation and Qwen2.5-7B fusion required 1.60 ± 0.58 s and 2.36 ± 0.29 s/video, respectively. InternVL3 and Qwen2.5-7B each occupied approximately 43 GB of GPU memory as resident vLLM servers on separate GPUs. Additional details on the profiling protocol, hardware configuration and stage-wise measurements are provided in the Supplementary material.

5.2 Qualitative examples

Figures 2, 3 complement the aggregate results with two representative DFEW examples, one in which the fusion model corrects the MIL prediction and one in which it confirms it. In both, InternVL3-8B is used for contextual evidence extraction and Qwen2.5-7B-Instruct for evidence fusion. In video ID 188 (Figure 2), the MIL model initially predicts *neutral*, whereas the fusion model revises the prediction to the correct label, *sad*. At the selected frame, AU17 and AU20 provide facial evidence compatible with sadness, and the VLM-generated contextual description reinforces it by reporting visible tears. Video ID 6 (Figure 3) shows the complementary case, in which the MIL model already predicts the correct class, *anger*. The selected frame contains AU11, AU17, and AU23 activations compatible with anger, and the VLM describes a tense face-to-face interaction. The fusion model keeps the original prediction and produces an explanation consistent with both the facial and the contextual evidence.

Together, these examples illustrate two possible behaviors of the fusion stage: revising an initial prediction when complementary evidence supports an alternative interpretation, and retaining an already correct prediction when the available evidence is mutually consistent. They are intended as qualitative illustrations

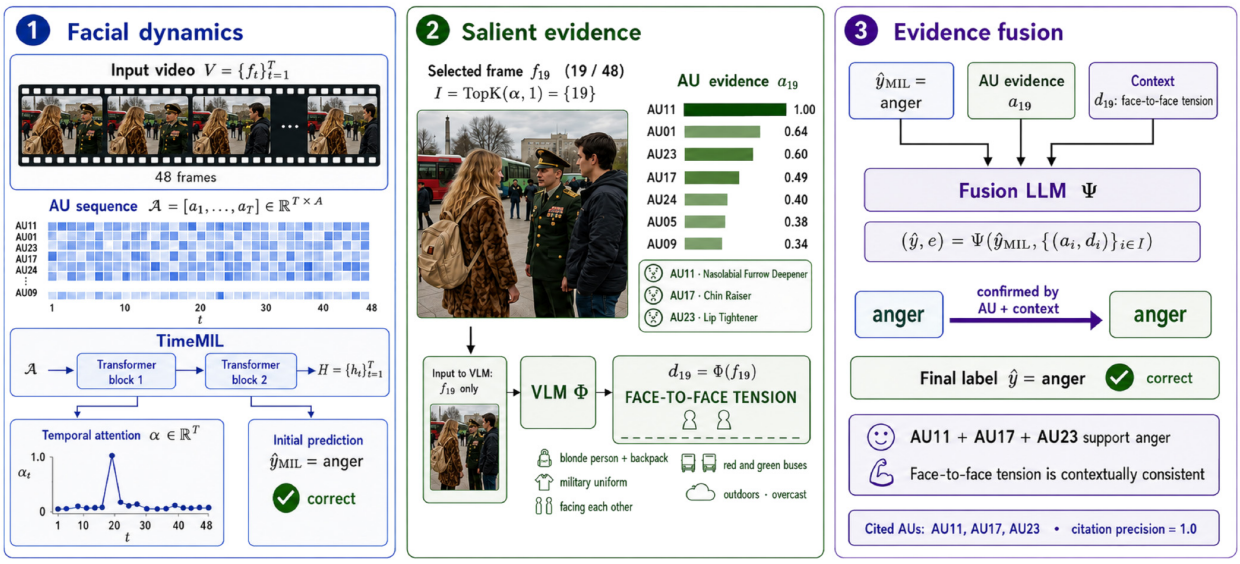

FIGURE 3
ReFER full-agreement example for video ID 6 from DFEW. MIL model correctly predicts anger and assigns the highest temporal relevance to frame 19/48. At the selected frame, AU11, AU17 and AU23 provide facial evidence supporting anger, while VLM identifies a potentially tense face-to-face interaction. LLM integrates the initial prediction, AU measurements, and contextual description, retaining the correct final label anger and generating an evidence-grounded explanation. The frames shown are regenerated visualizations rather than reproductions of the original DFEW footage, as permission permits use of the dataset but not reproduction or distribution of its data. Images were generated using OpenAI ChatGPT Images 2.0.

TABLE 5 Ablation study isolating prompt components (Emotion label, VLM Context) and Frame Source (`salient` vs. `random`).

Emotion	Context	Frame	AFEW WAR (%)	AFEW UAR (%)	DFEW WAR (%)	DFEW UAR (%)	MAFW WAR (%)	MAFW UAR (%)	FERV39k WAR (%)	FERV39k UAR (%)
✓	✓	Salient	**40.3**	**37.8**	53.5±1.4	43.3±1.3	33.18	22.72	**37.86**	**27.50**
		Random	39.8	37.3	**53.8±1.3**	**43.6±1.2**	**33.28**	**22.79**	37.26	27.30
✓	✗	Salient	**38.5**	**36.2**	**51.0±1.6**	**41.4±1.4**	29.48	20.43	**35.51**	**25.73**
		Random	36.9	34.8	50.7±1.2	41.2±1.2	**30.54**	**21.24**	34.58	25.24
✗	✓	Salient	**20.2**	**20.8**	25.6±1.3	23.8±1.0	20.09	**14.63**	**23.77**	21.39
		Random	19.4	19.5	**27.2±0.9**	**24.3±0.7**	**20.80**	14.54	23.72	**21.60**
✗	✗	Salient	**15.2**	**15.5**	22.1±1.4	18.6±1.0	15.65	12.09	**21.74**	**18.20**
		Random	14.9	14.4	**23.8±0.5**	**19.4±0.4**	**16.87**	**12.19**	21.35	17.94

Fusion backbone: Qwen2.5-7B-Instruct. DFEW values are mean over five-fold; AFEW and MAFW are single-split; FERV39k follows its official split. `random` values are averaged over three seeds. Bold marks the better of `salient`/`random` within each block, per dataset and metric.

of the reasoning process, while aggregate behavior is reported in Tables 2–4.

5.3 Ablations

5.3.1 Component analysis

We analyze the contribution of the main information sources provided to the fusion model by selectively removing the initial prediction ($\hat{y}_{\mathrm{MIL}}$) or the VLM-derived contextual description. The AU measurements are retained in all configurations, as they constitute the primary structured evidence used by ReFER. We additionally compare attention-guided (salient) frame selection against a baseline that selects the same number of frames uniformly at random, averaged over three seeds, to test whether MIL's attention scores identify more informative frames than arbitrary selection.

Table 5 shows that removing either the initial MIL prediction or the contextual description degrades performance, but the two components are not interchangeable: removing the emotion label (rows 3–4) causes a substantially larger drop than removing the context (row 2), reducing DFEW WAR from 53.5 to 25.6% with salient frames. Although salient selection provides modest numerical gains in some configurations, a paired bootstrap analysis does not identify a statistically reliable advantage over random selection on any of the four datasets. These results suggest that, in the current setting, MIL-guided frame selection does not appear

TABLE 6 Contextual corrections among samples misclassified by the MIL baseline.

Dataset	Baseline errors (*N*)	Contextual corrections (%)
AFEW	233	2.6 (1.2–5.5)
DFEW	5,564	3.3 (2.9–3.8)
MAFW	1,195	3.9 (3.0–5.2)
FERV39k	3,827	4.8 (4.1–5.5)

A contextual correction is a sample that remains incorrect in the No-Context configuration but becomes correct after adding the VLM-derived contextual description. Values in brackets report 95% Wilson confidence intervals.

TABLE 7 Comparison with lightweight fusion baselines on DFEW.

Method	WAR (%)	UAR (%)
MIL (baseline)	52.43 ± 1.28	42.17 ± 1.25
Logistic regression	49.19 ± 1.45	**43.71 ± 1.64**
Small MLP	48.94 ± 1.41	43.48 ± 1.82
ReFER No-Context	51.03 ± 1.57	41.42 ± 1.41
ReFER Full (Qwen2.5-7B)	**53.49 ± 1.38**	43.34 ± 1.34

Results are reported as mean ± standard deviation over the five official folds. Bold values indicate the best results.

to consistently improve downstream contextual reasoning over random sampling. Detailed confidence intervals are reported in the Supplementary material, and further discussion in Section 5.4.

5.3.2 Context contribution analysis

To more directly assess when contextual information improves the initial prediction, we focus on samples misclassified by the MIL baseline. We then compare the matched Full and No-Context configurations and count a *contextual correction* when a sample remains misclassified without context but is correctly classified after adding the VLM-derived contextual description. This isolates corrections specifically associated with contextual evidence, rather than improvements already obtained from AU-based fusion alone. As shown in Table 6, contextual information corrects a small but consistent subset of MIL errors across all four datasets, ranging from 2.6% on AFEW to 4.8% on FERV39k. These results indicate that context provides a selective contribution, becoming useful for a subset of cases in which the initial prediction and AU information alone are insufficient.

5.3.3 Fusion baselines

To assess whether ReFER's gains can be reproduced by simpler fusion mechanisms, we compare the proposed pipeline with two lightweight classifiers operating on TimeMIL output probabilities and AU measurements: logistic regression and a small MLP. We additionally report the ReFER No-Context configuration, which retains LLM-based fusion but removes the VLM-derived contextual description. As shown in Table 7, the lightweight baselines do not improve over TimeMIL in WAR, despite achieving comparable UAR, while the No-Context variant also remains below the Full pipeline. Overall, the strongest performance is obtained when LLM-based fusion is combined with contextual evidence, suggesting that neither shallow fusion of MIL and AU information nor LLM fusion without context is sufficient to reproduce the complete ReFER behavior.

5.3.4 Number of selected frames

We analyze the effect of the number of selected frames k on DFEW fold 1 using Qwen2.5-7B-Instruct. Table 8 shows that increasing k yields limited and non-monotonic gains. In the complete configuration, performance peaks at $k = 5$, improving over $k = 1$ by only 0.60 and 0.50 percentage points in WAR and UAR, respectively, and decreases for larger values. We therefore retain $k = 1$ as a practical operating point: it achieves performance close to the best observed while minimizing the number of frames processed by InternVL3 for contextual evidence extraction. Since the remaining pipeline components are unchanged with k, reducing the number of selected frames directly limits the additional VLM inference cost associated with contextual reasoning. With $k = 1$, contextual description generation requires approximately 1.60 s/video in our computational profile, making the number of VLM-processed frames the main variable cost controlled by k.

5.3.5 Label-anchoring analysis

To assess whether the fusion model simply reproduces the initial TimeMIL prediction rather than reasoning over the supplied evidence, we perform a label-corruption intervention on DFEW. For each video, we replace the original prediction $\hat{y}_{\text{MIL}}$ with a label sampled from the remaining emotion classes, while leaving the AU measurements and contextual description unchanged, and then run the fusion model as usual. We compute the *label-adoption rate*, defined as the percentage of samples for which the final prediction matches the corrupted (incorrect) label rather than the class supported by the AU and contextual evidence; higher values indicate greater susceptibility to naively following the emotion label provided in the prompt, i.e., "anchoring" on $\hat{y}_{\text{MIL}}$ rather than reasoning over the evidence. This analysis is conducted for all five fusion models over the five official DFEW folds.

As shown in Table 9, all fusion models exhibit strong susceptibility to the label supplied in the prompt. Label-adoption rates range from 82.60% for Llama-3.1-8B to 99.90% for Llama-2-7B, showing that even the most resistant model follows the corrupted prediction in the large majority of cases. These results indicate that the initial MIL prediction acts as a strong anchor for the fusion model, rather than merely as one source of evidence among several. Therefore, improvements from the full pipeline alone do not show that the LLM independently reconciles the AU and contextual observations.

5.3.6 Anchoring mitigation

Given the strong label anchoring observed above, we further investigate whether this behavior can be mitigated at inference

TABLE 8 Ablation study on the number of selected frames k across prompt component configurations (Emotion label, VLM Context).

		$k = 1$		$k = 3$		$k = 5$		$k = 7$		$k = 10$	
Emotion	Context	WAR	UAR	WAR	UAR	WAR	UAR	WAR	UAR	WAR	UAR
✓	✓	53.35	43.29	53.82	43.60	**53.95**	**43.79**	53.44	43.14	53.31	43.07
✓	✗	51.64	41.98	52.71	42.56	52.67	42.56	**52.93**	**42.69**	**52.93**	**42.69**
✗	✓	**26.91**	**24.33**	25.42	22.04	25.20	22.49	25.97	22.70	25.59	21.75
✗	✗	**23.96**	**19.50**	22.90	18.15	22.98	19.14	23.07	18.83	23.32	18.89

Fusion backbone: Qwen2.5-7B-Instruct. Results are reported on DFEW fold 1. Bold marks the best value across k within each configuration, per metric.

TABLE 9 Label-anchoring analysis on DFEW.

Metric	Qwen2.5-7B	Mistral-7B	Llama-3.1-8B	Llama-2-7B	EmoLlama-7B
Label-adoption rate (%)	97.9 ± 0.3	89.9 ± 0.6	82.6 ± 1.2	99.9 ± 0.1	99.6 ± 0.1

The label-adoption rate measures the percentage of samples for which the final prediction matches the corrupted $\hat{y}_{\text{MIL}}$ label. Results are reported as mean ± standard deviation over the five official folds. Lower values indicate greater resistance to the provided label.

TABLE 10 Inference-time strategies for mitigating anchoring on DFEW with Qwen2.5-7B-Instruct.

Method	WAR (%)	UAR (%)	Label adoption (%)
ReFER Full	53.50 ± 1.37	43.34 ± 1.33	97.86 ± 0.27
Prediction-blind	25.59 ± 1.29	23.82 ± 1.00	–
Two-stage	25.59 ± 1.29	23.82 ± 1.00	–
Symmetric reconciliation	33.89 ± 1.12	30.98 ± 0.82	41.21 ± 1.09

Corrupted-label adoption measures how often the final prediction matches the corrupted MIL label.

time without modifying the frozen fusion model. As a reference condition, we consider the *prediction-blind* configuration, in which the fusion LLM receives AU and contextual evidence but not the initial MIL prediction. We first consider a two-stage strategy in which the LLM forms an evidence-only prediction from AU and contextual information before being exposed to the MIL prediction for reconciliation. We additionally test a symmetric reconciliation strategy, in which the MIL-conditioned and prediction-blind predictions are presented as two anonymous candidate hypotheses together with the original evidence. The exact prompt templates for both strategies are reported in the Supplementary material. As shown in Table 10, both strategies reduce corrupted-label adoption, but at a substantial cost in recognition performance. The two-stage setting collapses to the prediction-blind behavior, as the reconciliation step systematically retains the evidence-only prediction and never overrides it in favor of the MIL prediction. Symmetric reconciliation provides a less extreme trade-off, but still remains below standard ReFER in terms of WAR and UAR. Moreover, swapping the order of the two anonymous candidates changes the final prediction in 38.9% of disagreement cases, revealing a substantial position bias. Overall, inference-time prompting alone does not resolve the tension between recognition performance and reliance on the initial MIL prediction.

5.4 Limitations

First, attention-guided frame selection does not show a statistically reliable advantage over random selection on any of the evaluated datasets. On DFEW, the largest benchmark, random selection slightly but consistently outperforms the attention-guided strategy. This result does not necessarily imply that the learned attention is uninformative; rather, it highlights a mismatch between the objective used to learn the attention weights and the purpose for which the selected frames are subsequently used. MIL attention is optimized to identify temporal instances that support video-level emotion classification from AU dynamics, whereas the VLM benefits from frames containing complementary contextual evidence, such as posture, interactions, objects, or scene composition. A frame that is discriminative for the MIL classifier is therefore not necessarily the frame that is most informative for contextual interpretation. Moreover, these benchmarks predominantly contain short clips, often spanning only a few seconds, so contextual information may be highly redundant across frames. This may reduce the potential benefit of selecting a specific frame over random sampling. In longer and more temporally heterogeneous videos, where relevant contextual cues may appear only during limited portions of the sequence, frame selection could play a more important role; evaluating this hypothesis remains an interesting direction for future work. MIL attention should consequently be regarded as an operational mechanism for evidence localization, rather than as a validated measure of contextual informativeness or human-interpretable saliency.

Second, the label-corruption analysis reveals substantial anchoring to the initial MIL prediction. When an artificially corrupted MIL label influences the final decision despite the availability of AU and contextual evidence, the fusion LLM behaves less as an independent evidence integrator and more as a label-conditioned verifier or reviser. The mitigation experiments further show that this dependence cannot be straightforwardly resolved through inference-time prompting

without sacrificing recognition performance or introducing other biases. Consequently, a coherent explanation does not by itself demonstrate that the model reconciled the evidence faithfully: the LLM may construct a plausible *post-hoc* justification around the supplied prediction. The deterministic grounding metrics and LLM-based judges used in this work assess useful properties of the generated text, including AU citation accuracy, contextual consistency, and prediction coherence, but cannot establish causal faithfulness to the model's underlying decision process.

Finally, the pipeline inherits the errors and biases of its frozen upstream components, including unreliable AU estimates under unconstrained pose, illumination, or occlusion and VLM descriptions that may omit, distort, or misrepresent relevant visual evidence.

6 Conclusion

We presented ReFER, a framework that extends AU-based temporal emotion recognition by combining an initial MIL prediction with localized facial measurements and VLM-derived contextual descriptions through an instruction-tuned fusion LLM. The framework produces both a final emotion prediction and a natural-language explanation grounded in the supplied evidence. We evaluated five fusion models across four in-the-wild video FER benchmarks, considering both recognition performance and complementary measures of explanation quality.

Our results show that LLM-based fusion can improve the AU-only baseline, but that these gains depend strongly on the fusion model and dataset. Contextual evidence provides complementary information, yet the initial MIL prediction remains the dominant input to the fusion stage. At the same time, the analyses reveal that explanation coherence and evidence grounding should not be interpreted as evidence of faithful decision-making, since the fusion models remain strongly susceptible to the prediction supplied in the prompt. Overall, these findings highlight both the potential and the current limitations of language-model-based evidence fusion for video emotion recognition.

Taken together, ReFER provides a systematic setting for studying how structured facial evidence, visual context and language models interact in video emotion recognition. Future work should investigate adaptive context aggregation and frame-selection objectives designed to retrieve information complementary to the MIL representation. Training strategies based on masking or deliberately corrupting the initial prediction, or fusion mechanisms explicitly designed to handle conflicting evidence, may provide a more robust way to reduce anchoring. Beyond LLM-based prompting, future work could also investigate other simple fusion mechanisms, such as calibrated decision-level fusion or lightweight gating, which may provide more explicit control over the relative contribution of the MIL prediction and complementary evidence. Finally, counterfactual evaluation and human assessment of the factual correctness, relevance and faithfulness of the generated explanations will be important for determining whether future systems genuinely reconcile their evidence rather than construct plausible *post-hoc* justifications.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

LD: Software, Writing – original draft, Investigation, Formal analysis, Conceptualization, Visualization, Methodology. VC: Methodology, Writing – review & editing, Conceptualization, Validation, Supervision. RC: Writing – review & editing, Supervision, Project administration, Resources.

Funding

The author(s) declared that financial support was received for this work and/or its publication. This work is partially funded by IRCCS Istituto delle Scienze Neurologiche di Bologna (IRCCS-ISNB) and Takeda Development Center Americas, Inc. Takeda Development Center Americas, Inc. was not involved in the study design, collection, analysis, interpretation of data, the writing of this article, or the decision to submit it for publication.

Acknowledgments

We acknowledge ISCRA for awarding this project access to the LEONARDO supercomputer, owned by the EuroHPC Joint Undertaking, hosted by CINECA.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that Generative AI was used in the creation of this manuscript. Generative AI was used as a writing support tool and figures generator during the preparation of the manuscript, primarily to assist with language refinement, rephrasing, and improving the clarity and readability of the text. The scientific content, methodological choices, experimental results, analyses, and conclusions were developed and verified by the author(s). The images in Figures 1–3 were generated using OpenAI ChatGPT Images 2.0.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2026.1984257/full#supplementary-material

References

Aviezer, H., Bentin, S., Dudarev, V., and Hassin, R. R. (2008). Angry, disgusted, or afraid? Studies on the malleability of emotion perception. *Psychol. Sci.* 19, 724–732. doi: 10.1111/j.1467-9280.2008.02148.x

Aviezer, H., Trope, Y., and Todorov, A. (2012). Body cues, not facial expressions, discriminate between intense positive and negative emotions. *Science* 338, 1225–1229. doi: 10.1126/science.1224313

Barrett, L. F., Mesquita, B., and Gendron, M. (2011). Context in emotion perception. *Curr. Dir. Psychol. Sci.* 20, 286–290. doi: 10.1177/0963721411422522

Bhattacharya, P., Gupta, R. K., and Yang, Y. (2021). "Exploring the contextual factors affecting multimodal emotion recognition in videos," in *IEEE transactions on affective computing, Vol. 14*, 1547–1557. doi: 10.1109/TAFFC.2021.3071503

Boccignone, G., Cuculo, V., D'Amelio, A., Grossi, G., and Lanzarotti, R. (2020). On gaze deployment to audio-visual cues of social interactions. *IEEE Access* 8, 161630–161654. doi: 10.1109/ACCESS.2020.3021211

Carroll, J. M., and Russell, J. A. (1996). Do facial expressions signal specific emotions? Judging emotion from the face in context. *J. Pers. Soc. Psychol.* 70, 205–218. doi: 10.1037/0022-3514.70.2.205

Chen, X., Qiu, P., Zhu, W., Li, H., Wang, H., Sotiras, A., et al. (2024). TimeMIL: advancing multivariate time series classification via a time-aware multiple instance learning. *arXiv*. [preprint]. arXiv:2405.03140. doi: 10.48550/arXiv.2405.03140

Cheong, J. H., Jolly, E., Xie, T., Byrne, S., Kenney, M., and Chang, L. J. (2023). Py-Feat: python facial expression analysis toolbox. *Affect. Sci.* 4, 781–796. doi: 10.1007/s42761-023-00191-4

Colaco, S. J., and Han, D. S. (2025). Scalable context-based facial emotion recognition using facial landmarks and attention mechanism. *IEEE Access* 13, 20778–20791. doi: 10.1109/ACCESS.2025.3534328

Del Gaudio, L., Cuculo, V., and Cucchiara, R. (2026a). "Decoding facial expressions in video: a multiple instance learning perspective on action units," in *Image analysis and processing-ICIAP 2025 workshops, lecture notes in computer science* (Cham: Springer).

Del Gaudio, L., Cuculo, V., and Cucchiara, R. (2026b). "HyperMIL: hypergraph-based channel reasoning for multiple instance learning on multivariate time series," in *International conference on pattern recognition* (Cham: Springer), 208–222.

Dhall, A., Goecke, R., Joshi, J., Sikka, K., and Gedeon, T. (2014). "Emotion recognition in the wild challenge 2014: baseline, data and protocol," in *Proceedings of the 16th international conference on multimodal interaction* (Istanbul: ACM), 461–466. doi: 10.1145/2663204.2666275

Ekman, P., and Friesen, W. V. (1978). *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Palo Alto, CA: Consulting Psychologists Press. doi: 10.1037/t27734-000

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., et al. (2024). The Llama 3 herd of models. *arXiv*. [preprint]. arXiv:2407.21783. doi: 10.48550/arXiv.2407.21783

Hoang, M.-H., Kim, S.-H., Yang, H.-J., and Lee, G.-S. (2021). Context-aware emotion recognition based on visual relationship detection. *IEEE Access*, 9, 90465–90474. doi: 10.1109/ACCESS.2021.3091169

Huang, J., Lin, F., Yang, X., Feng, C., Zhu, K., Yang, X., et al. (2026). "Nano-EmoX: unifying multimodal emotional intelligence from perception to empathy," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pages* 22986-22997.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D., et al. (2023). Mistral 7b. *arXiv*. [preprint]. arXiv:2310.06825. doi: 10.48550/arXiv.2310.06825

Jiang, X., Zong, Y., Zheng, W., Tang, C., Xia, W., Lu, C., et al. (2020). "DFEW: a large-scale database for recognizing dynamic facial expressions in the wild," in *Proceedings of the 28th ACM international conference on multimedia* (New York, NY: ACM), 2881–2889. doi: 10.1145/3394171.3413620

Kosti, R., Alvarez, J. M., Recasens, A., and Lapedriza, A. (2017). "Emotion recognition in context," in *IEEE conference on computer vision and pattern recognition (CVPR)* (Honolulu, HI: IEEE), 1667–1675. doi: 10.1109/CVPR.2017.212

Lee, J., Kim, S., Kim, S., Park, J., and Sohn, K. (2019). "Context-aware emotion recognition networks," in *Proceedings of the IEEE/CVF international conference on computer vision* (Piscataway, NJ: IEEE), 10143–10152. doi: 10.1109/ICCV.2019.01024

Lian, Z., Sun, H., Sun, L., Gu, H., Wen, Z., Zhang, S., et al. (2023). "Explainable multimodal emotion recognition," *arXiv*. [preprint]. arXiv:2306.15401. doi: 10.48550/arXiv.2306.15401

Liu, Y., Dai, W., Feng, C., Wang, W., Yin, G., Zeng, J., et al. (2022). "MAFW: a large-scale, multi-modal, compound affective database for dynamic facial expression recognition in the wild," in *ACM MM* (New York, NY: ACM), 24–32. doi: 10.1145/3503161.3548190

Liu, Z., Yang, K., Xie, Q., Zhang, T., and Ananiadou, S. (2024). "EmoLLMs: a series of emotional large language models and annotation tools for comprehensive affective analysis," in *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining* (New York, NY: ACM), 5487–5496. doi: 10.1145/3637528.3671552

Mittal, T., Bera, A., and Manocha, D. (2021). "Multimodal and context-aware emotion perception model with multiplicative fusion," in *IEEE multimedia* (Piscataway, NJ: IEEE). doi: 10.1109/MMUL.2021.3068387

Mittal, T., Guhan, P., Bhattacharya, U., Chandra, R., Bera, A., and Manocha, D. (2020). "EmotiCon: context-aware multimodal emotion recognition using Frege's principle," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14234–14243.

Rha, H., Yeo, J. H., Kim, Y., and Ro, Y. M. (2026). "Emotion-coherent reasoning for multimodal LLMs via emotional rationale verifier," in *Proceedings of the AAAI conference on artificial intelligence* (Washington, DC: AAAI Press), number 3, 2029–2037.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., et al. (2023). Llama 2: open foundation and fine-tuned chat models. *arXiv*. [preprint]. arXiv:2307.09288. doi: 10.48550/arXiv.2307.09288

Wang, Y., Sun, Y., Huang, Y., Liu, Z., Gao, S., Zhang, W., et al. (2022). "FERV39k: a large-scale multi-scene dataset for facial expression recognition in videos," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (New Orleans, LA: IEEE), 20922–20931. doi: 10.1109/CVPR52688.2022.02025

Wang, Z., Lao, L., Zhang, X., Li, Y., Zhang, T., and Cui, Z. (2022). Context-dependent emotion recognition. *J. Vis. Commun. Image Represent.* 89: 103679. doi: 10.1016/j.jvcir.2022.103679

Wieser, M. J., and Brosch, T. (2014). Faces in context: a review and systematization of contextual influences on affective face processing. *Front. Psychol.* 3:471. doi: 10.3389/fpsyg.2012.00471

Wu, J., Shen, Y., Yan, L., Sun, H., Xia, D., Huang, J., et al. (2026). "Facial-R1: aligning reasoning and recognition for facial emotion analysis," in *Proceedings of the AAAI conference on artificial intelligence* (Washington, DC: AAAI Press), Vol. 40, 26939–26947.

Wu, S., Zhou, L., Hu, Z., and Liu, J. (2022). Hierarchical context-based emotion recognition with scene graphs. *IEEE transactions on neural networks and learning systems* (IEEE), 35, 3725–3739. doi: 10.1109/TNNLS.2022.3196831

Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., et al. (2024). Qwen2.5 technical report. *arXiv*. [preprint]. arXiv:2412.15115. doi: 10.48550/arXiv.2412.15115

Yang, D., Huang, S., Wang, S., Liu, Y., Zhai, P., Su, L., et al. (2022). "Emotion recognition for multiple context awareness," in *European conference on computer vision* (Cham: Springer), 144–162.

Zhang, H., Liu, Y., Jiang, P., Junjie, L., Jun, X., He, Y., et al. (2026). XEmoGPT: an explainable multimodal emotion recognition framework with cue-level perception and reasoning. *arXiv*. [preprint]. arXiv:2602.05496. doi: 10.48550/arXiv.2602.05496

Zhang, L., Zheng, X., Chen, X., and Cui, L. (2024). "Three-dimensional view relationship-based context-aware emotion recognition," in *IEEE Trans Neural Netw Learn Syst, Vol. 36* (IEEE), 13567–13578. doi: 10.1109/TNNLS.2024.3476249

Zhang, Z., Wang, W., Zhu, Y., Qin, W., Wan, P., Zhang, D., et al. (2026). VidEmo: affective-tree reasoning for emotion-centric video foundation models. *Adv. Neural Inf. Process. Syst*. 38, 80843–80890. doi: 10.52202/085713-2703

Zhao, J., Wei, X., and Bo, L. (2025). R1-Omni: explainable omni-multimodal emotion recognition with reinforcement learning. *arXiv*. [preprint]. arXiv:2503.05379. doi: 10.48550/arXiv.2503.05379

Zhao, Z., Tian, W., Tian, X., Zhang, J., and Xie, L. (2026). Integrating fine-grained audio-visual evidence for robust multimodal emotion reasoning. *arXiv*. [preprint]. arXiv:2601.18321. doi: 10.48550/arXiv.2601.18321

Zheng, K., Fu, J., Deng, W., Han, H., Jose, J. M., and Ge, X. (2026). Do we really need multimodal emotion language models larger than 1B parameters? *arXiv*. [preprint]. arXiv:2607.12787. doi: 10.48550/arXiv.2607.12787

Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., et al. (2025). InternVL3: exploring advanced training and test-time recipes for open-source multimodal models. *arXiv*. [preprint]. arXiv:2504.10479. doi: 10.48550/arXiv.2504.10479