CERA: Conflict-Explicit Reflective Agent for Multimodal Emotion Reasoning

Kejun Liu
China University of Geosciences
Wuhan, China
liukejun@cug.edu.cn

Yuanyuan Liu*
China University of Geosciences
Wuhan, China
liuyy@cug.edu.cn

Ke Wang
China University of Geosciences
Wuhan, China
WK2023@cug.edu.cn

Jiahao Zhang
China University of Geosciences
Wuhan, China
zhangjiahao2@cug.edu.cn

Lei Xu
China University of Geosciences
Wuhan, China
1202411237@cug.edu.cn

Chang Tang
Huazhong University of Science and Technology
Wuhan, China
tangchang@hust.edu.cn

Zhe Chen
La Trobe University
Melbourne, Australia
zhe.chen@latrobe.edu.au

Yibing Zhan
Wuhan University
Wuhan, China
zybjy@mail.ustc.edu.cn

Abstract

Multimodal Emotion Recognition (MER) aims to understand complex human emotions by jointly analyzing visual and textual data. However, in real-world scenarios, emotional cues from different modalities often contain conflict information, such as a smiling face paired with negative text, which poses great challenges for existing multimodal language models (MLLMs). Existing emotion MLLMs and multimodal emotion benchmarks often overlook or even intentionally avoid scenarios involving multimodal emotion conflicts, limiting their ability to reason about complex and contradictory affective cues. By addressing this, we propose Conflict-Explicit Reflective Agent (CERA), a training-free, conflict-aware, and language-driven agentic framework for MER. The concept of CERA is to treat modality emotion conflicts as meaningful signals and resolve them via a three-stage perception–evaluation–reflection reasoning loop. Firstly, the agent's conflict-perceptive emotion graph construction module builds emotion graphs from fine-grained cues to reveal conflicts, and progressively refines them through iterative updates. Secondly, a reward model evaluates these graphs and produces natural language feedback that identifies unresolved conflicts. Lastly, the language-driven conflict refinement module generates graph editing signals from the feedback without any parameter tuning, enabling the overall CERA to refine its reasoning without training. Extensive experiments on two multimodal emotion datasets, MAFW and CH-SIMS, demonstrate that CERA significantly outperforms state-of-the-art training-free methods in both recognition accuracy and conflict interpretability, providing an effective training-free solution for complex emotional reasoning.

CCS Concepts

• **Computing methodologies** → **Artificial intelligence**.

Keywords

multimodal emotion recognition, cross-modal conflict, emotion graph, training-free inference, language-driven refinement

ACM Reference Format:
Kejun Liu, Yuanyuan Liu, Ke Wang, Jiahao Zhang, Lei Xu, Chang Tang, Zhe Chen, and Yibing Zhan. 2026. CERA: Conflict-Explicit Reflective Agent for Multimodal Emotion Reasoning. In *International Conference on Multimedia Retrieval (ICMR '26), June 16–19, 2026, Amsterdam, Netherlands.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3805622.3810634

1 Introduction

Multimodal Emotion Recognition (MER) aims to model complex human emotions from multimodal signals such as video and text, constituting a core problem in affective computing and a high-level semantic signal for emotion-aware multimedia retrieval and human-centered content understanding [8, 17, 26]. Recent advances in Large Multimodal Language Models (MLLMs) have significantly improved multimodal understanding, offering new opportunities for emotion-related tasks. With strong cross-modal reasoning and emotionally aware generation capabilities, MLLMs have shown promise in modeling nuanced affective states. For example, Emotion-LLaMA [6] integrates emotion-specific encoders and instruction tuning to improve affective understanding in complex scenarios. Recent training-free approaches, including chain-of-thought (CoT) prompting [38], indicate that MLLM can perform affective reasoning without task-specific fine-tuning, offering advantages in efficiency and interpretability.

Despite these advances, current MLLMs face a fundamental challenge: ***modality emotion conflicts***, which refer to inconsistencies

*Corresponding authors.

ICMR '26, Amsterdam, Netherlands

ACM ISBN 979-8-4007-2617-0/2026/06
https://doi.org/10.1145/3805622.3810634

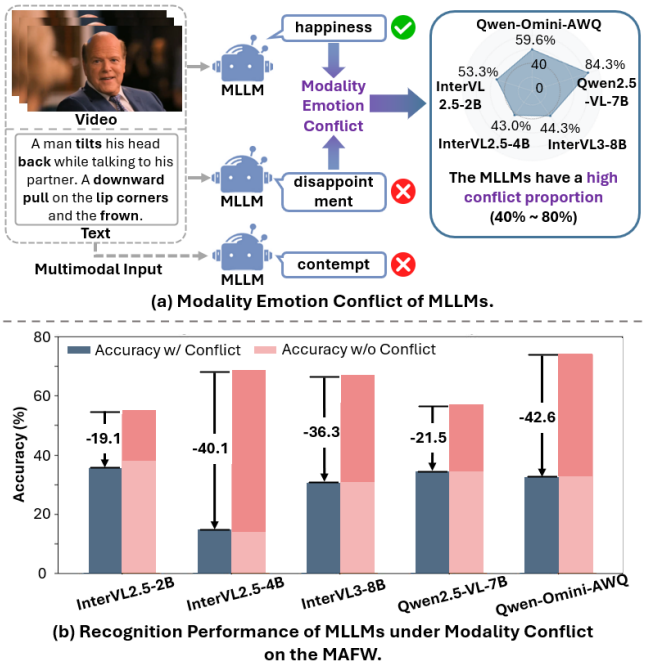

Figure 1: Analysis of Modality Emotion Conflicts in MLLMs on the MAFW Dataset. (a)Illustration of modality emotion conflicts where MLLMs produce inconsistent predictions across modalities. Most MLLMs exhibit high conflict proportion (40%–80%). (b) Accuracy comparison in conflict vs. no conflict scenarios. All models suffer notable drops (19.1%–42.6%) under conflict, revealing limited ability to handle emotional inconsistency.

between emotional cues across modalities. As shown in Figure 1, while the visual input conveys a pleased expression, the textual description highlights negative signals such as "frown" and "downward pull," creating a modality emotion conflict that challenges MLLMs' reasoning. However, existing emotion MLLMs and multimodal emotion benchmarks often overlook or even intentionally avoid such conflict scenarios [15, 30], limiting their capacity to reason about complex and contradictory affective cues. For instance, Omni-Emotion [30] explicitly discards emotionally inconsistent samples to maintain dataset purity, thereby preventing models from learning to reason under cross-modal contradictions. Recently, MoSEAR [9] took an important step forward by benchmarking and bridging emotion conflicts for multimodal emotion reasoning. Nevertheless, MoSEAR mainly alleviates inconsistencies through attention redistribution and contrastive regularization, rather than explicitly modeling or utilizing conflicts as informative cues. In contrast, we argue that modality emotion conflicts convey subtle but important signals, such as emotion suppression, deception, or concealed affect, all of which are essential for accurately understanding a person's true emotional state. Ignoring these cues not only limits a model's ability to reason about human emotion but also undermines interpretability and alignment with human cognitive processes.

To quantify the impact of modality emotion conflicts, we conducted a comprehensive analysis on the MAFW dataset using five representative MLLMs, including InterVL2.5-2B/4B [4], InterVL3-8B [40], Qwen2.5-VL [2], and Qwen-Omni-AWQ [29]. As shown in Figure 1, such modality emotion conflicts are pervasive, with conflict proportion[1] ranging from 43% to 84.2%. Emotion recognition accuracy drops sharply (by 19.1% to 42.6%) under conflict scenarios, and larger model sizes do not alleviate the issue. These findings reveal an important gap: existing MLLMs lack structured mechanisms to perceive and reason over emotional inconsistencies.

To address this, we propose Conflict-Explicit Reflective Agent (CERA), a training-free, conflict-aware, and language-driven agentic framework for MER. Unlike existing models that suppress conflicting signals, CERA is designed to explicitly perceive, evaluate, and resolve such conflicts as informative cues of suppressed or concealed emotions. The CERA achieves this goal via a structured perception–evaluation–reflection reasoning loop that enables the agent to iteratively refine its emotional understanding without any parameter updates: (1) Conflict-Perceptive Emotion Graph Construction (CEGC) builds and iteratively refines intra- and inter-modal emotion graphs that make emotional inconsistencies explicit; (2) Reward-Guided Conflict Feedback (RCF) applies a frozen reward model to evaluate the emotion graphs and generate natural language feedback on unresolved conflicts; (3) Language-Driven Conflict Refinement (LCR) converts this feedback into graph-level updates without any parameter tuning, allowing CERA to progressively refine its emotional reasoning. Here, the term "agent" denotes the entire reasoning pipeline that integrates a frozen MLLM, a reward model, and symbolic emotion graphs to reason over complex emotional dynamics. By modeling and refining modality emotion conflicts rather than suppressing them, CERA directly addresses the core challenge in MER: capturing nuanced emotional states that emerge through cross-modal inconsistencies. **Our contributions are summarized as follows:**

- We introduce **Conflict-Explicit Reflective Agent (CERA)**, a training-free, conflict-aware, and language-driven framework that treats modality emotion conflicts as informative and interpretable signals, and exploits them to improve robustness in multimodal emotion reasoning.
- We propose **Conflict-Perceptive Emotion Graph Construction (CEGC)**, which initializes and iteratively updates intra- and inter-modal emotion graphs to capture fine-grained emotional cues and explicitly encode cross-modal inconsistencies.
- We design a structured reasoning loop by integrating **Reward-Guided Conflict Feedback (RCF)** and **Language-Driven Conflict Refinement (LCR)**, enabling progressive conflict resolution through natural language feedback without any parameter tuning.
- We conduct comprehensive experiments on two benchmark datasets, MAFW and CH-SIMS, demonstrating that CERA consistently outperforms existing methods in both recognition accuracy and conflict interpretability.

2 Related Work

Multimodal Emotion Recognition. Multimodal Emotion Recognition (MER) seeks to understand human emotions by integrating

[1]The conflict proportion is defined as the percentage of samples for which the predicted emotion labels from the visual and textual modalities disagree.

signals from video, text, and audio. Early approaches employed feature-level fusion [1], but suffered from modality heterogeneity due to disparate semantic spaces. Attention-based methods [7] improved robustness by weighting salient cues but relied on manually crafted extractors. Decision-level fusion [19] aggregates unimodal outputs but misses fine-grained cross-modal interactions essential for decoding complex emotions like ambivalence or sarcasm. Despite significant progress, these methods continue to face challenges in modeling emotional transitions, addressing modality imbalance, and, most critically, handling modality emotion conflicts where emotional cues from different modalities are inconsistent.

MLLMs for Multimodal Emotion Understanding Large Multimodal Language Models (MLLMs) have advanced MER by enabling unified, instruction-following reasoning across modalities via shared language representations. Emotion-LLaMA [6] designed some emotion-specific encoders and employed instruction tuning to align and fuse multimodal features, significantly enhancing the ability of emotion recognition and reasoning. AffectGPT [14] constructed the largest fine-grained descriptive emotion dataset MER-Caption and a unified benchmark MER-UniBench specifically designed for evaluating the MER tasks of MLLMs. R1-Omni [39] applied Reinforcement Learning with Verifiable Reward (RLVR) to multimodal emotion understanding tasks for the first time, significantly enhancing the generalization ability, reasoning ability and interpretability of MLLMs in out-of-distribution scenarios.Despite strong understanding, MLLMs often fail under modality emotion conflicts, favoring coherent outputs over contradictory cues. This reveals a lack of structured reasoning critical for robust emotion understanding.

Handling Modality Emotion Conflicts in MER. Modality emotion conflicts occur when emotional cues from different modalities diverge, such as a smiling face paired with a sad caption. These conflicts are prevalent in real-world contexts and often convey hidden psychological meanings like suppression or deception. However, most existing MER methods either overlook or implicitly avoid such inconsistencies, focusing instead on enforcing cross-modal alignment. Early multimodal frameworks, including contrastive models such as CLIP [21] and ALBEF [12], minimize cross-modal discrepancies, which can inadvertently remove subtle emotional contrasts. Attention-based fusion methods [10, 35] reweight less dominant modalities under the assumption that inconsistencies hinder recognition. Other studies improve unimodal representations before fusion, such as the disentanglement-based MISA [27] and the self-supervised Self-MM [33], in order to preserve modality-specific semantics. Although these approaches enhance feature completeness and fusion robustness, they still pursue alignment as the final objective while overlooking the emotional implications of inter-modality discrepancies. Recently, several works have revisited modality emotion conflicts primarily as factors to be mitigated for more stable recognition. For example, MoSEAR [9] benchmarks conflict phenomena and alleviates them through redistribution strategies. Despite these advances, existing solutions rarely treat modality emotion conflicts themselves as informative emotional signals. This gap calls for models that explicitly perceive, evaluate, and reason about these conflicts as informative signals for more robust and interpretable emotion recognition.

3 Method

3.1 Pipeline Overview

Conflict-Explicit Reflective Agent (CERA) is a training-free, conflict-aware, and language-driven agentic framework designed to simulate human-like emotional reasoning. It operates via a structured perception–evaluation–reflection reasoning loop that treats modality emotion conflicts as meaningful cues for deeper emotional understanding. Although CERA relies on a multimodal large language model as its backbone, the "language" here does not introduce a new modality but serves as an intermediate interface for interpretable reflection and feedback in a training-free manner. The core stages include: **Conflict-perceptive Emotion Graph Construction (CEGC)**, **Reward-guided Conflict Feedback (RCF)**, and **Language-driven Conflict Refinement (LCR)** (Figure 2). Given an input video V and its corresponding text T, CERA first invokes **CEGC** to construct initial unimodal (G_v^{k}, G_t^{k}) and cross-modal (G_m^{k}) emotion graphs that capture fine-grained intra- and inter-modal emotional relationships. At each subsequent iteration $k \geq 1$, CEGC refines these graphs based on editing signals from LCR, progressively enhancing emotional structure and resolving modality emotion conflicts. Next, **RCF** employs a frozen reward model to evaluate the emotion graphs and prompts the MLLM to produce the language-level evaluation signal T_{eval}^{k} that highlights unresolved modality emotion conflicts and structural deficiencies. Finally, **LCR** transforms T_{eval} into language-guided graph editing signals ($g_v^{\mathrm{k}}, g_t^{\mathrm{k}}, g_m^{\mathrm{k}}$), providing interpretable update suggestions to guide CEGC in the next iteration of graph revision. This process simulates reflective reasoning through symbolic feedback without any parameter updates. This perception–evaluation–reflection loop is repeated for $K = 2$ iterations, progressively resolving modality emotion conflicts and enhancing emotional reasoning.

The refined final emotion graphs, combined with the original multimodal inputs, are used for emotion prediction, enabling CERA to achieve strong recognition performance and interpretability under challenging multimodal scenarios.

3.2 Conflict-perceptive Emotion Graph Construction (CEGC)

The goal of CEGC is to construct and refine structured emotion graphs that capture intra- and inter-modal emotional relationships, enabling explicit modeling of modality emotion conflicts. While existing MLLMs often summarize emotions at a global level [5], they tend to overlook localized emotional cues and subtle cross-modal inconsistencies critical for fine-grained understanding. To address this limitation, we introduce **emotion graphs** as a structured representation, where nodes represent fine-grained emotional cues (e.g., body posture, facial AU dynamics), and edges encode their emotional associations. This structure enables explicit modeling of intra- and inter-modal interactions, facilitating the detection of inconsistencies, redundancies, and complementary signals. By capturing these relational patterns, emotion graphs reveal latent conflicts that may indicate deeper psychological states such as suppression, contradiction, or concealed affect. Based on this design, CEGC constructs three types of graphs: the video emotion graph G_v^k, the text emotion graph G_t^k, and the cross-modal emotion graph

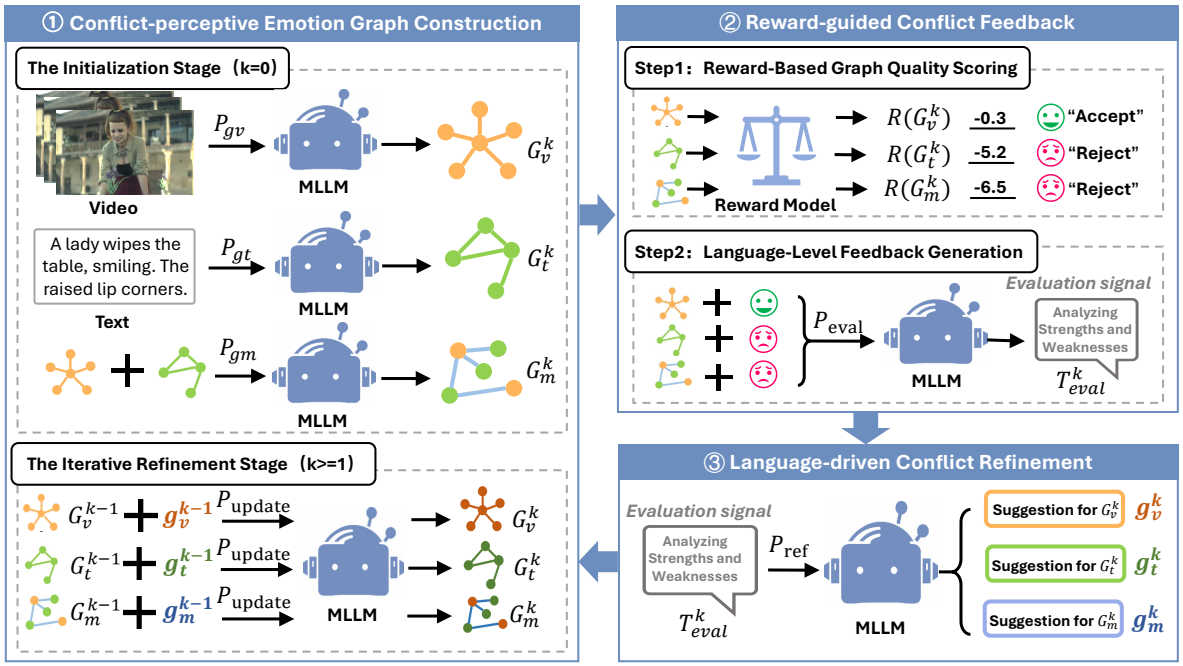

Figure 2: The overall pipeline of our proposed Conflict-Explicit Reflective Agent (CERA). (1) CEGC constructs and refines emotion graphs from video and text inputs; (2) RCF evaluates the graphs and generates language-level feedback; (3) LCR produces symbolic editing instructions that guide the next iteration of graph refinement. The final emotion graphs, together with the original inputs, are used for emotion prediction.

G_m^k. Within the iterative CERA loop, CEGC operates in two modes depending on the iteration step k: **the initialization stage** when $k = 0$, and **the iterative refinement stage** when $k \geq 1$.

3.2.1 The Initialization Stage ($k = 0$). Given a video clip V and its corresponding text T, we first prompt a frozen MLLM (denoted as M) with a video graph prompt P_{gv} to extract emotional cues from V, including facial expressions, body posture, and facial AU dynamics. ***Video emotion graph definition:*** these video cues are represented as nodes in the form "V:..." and edges are built between co-expressed cues that jointly imply one or more emotions. This yields the initial video emotion graph $G_v^k = \{\text{Node}_v, \text{Edge}_v\}$, where each edge follows the format: {"Head": "V:...", "Emotion": [...], "Tail": "V:..."}. The process is defined as:

$$G_v^k = M\left(P_{gv}\left(V\right)\right). \tag{1}$$

Similarly, we use a text graph prompt P_{gt} to extract emotional cues from T, including emotion words, tone indicators, and rhetorical indicators. ***Text emotion graph definition:*** these text cues are organized into the initial text emotion graph $G_t^k = \{\text{Node}_t, \text{Edge}_t\}$, where nodes are formatted as "T:...", and edges encode the jointly implied emotions between textual cues, following the structure: {"Head": "T:...", "Emotion": [...], "Tail": "T:..."}. Mathematically, this operation can be described as:

$$G_t^k = M\left(P_{gt}\left(T\right)\right). \tag{2}$$

To explicitly model modality emotion conflicts between modalities, we construct the initial cross-modal emotion graph G_m^k by feeding G_v^k and G_t^k into the MLLM under a cross-modal graph prompt P_{gm}. ***Cross-modal emotion graph definition:*** the resulting $G_m^k = \{\text{Node}_m, \text{Edge}_m\}$ connects nodes from $\text{Node}_v \cup \text{Node}_t$, and edges reflect the emotional relationships across video and text, following the structure:{"Head": "V:...", "Emotion": [...], "Tail": "T:..."}. The G_m^k is defined as:

$$G_m^k = M\left(P_{gm}\left(G_v^k, G_t^k\right)\right). \tag{3}$$

3.2.2 The Iterative Refinement Stage ($k \geq 1$). In subsequent iterations, CEGC updates G_v^{k-1}, G_t^{k-1}, and G_m^{k-1} based on the language-guided graph editing signals $(g_v^{k-1}, g_t^{k-1}, g_m^{k-1})$ generated by the LCR. Specifically, given $(g_v^{k-1}, g_t^{k-1}, g_m^{k-1})$, we construct an update prompt P_{update} to instruct the MLLM to revise the corresponding graphs. The editing operations include the removal of redundant or conflicting edges, the refinement of emotional labels, and the insertion of missing cues that enhance emotional completeness. The updated emotion graphs are obtained as:

$$G_v^k = M(P_{\text{update}}(G_v^{k-1}, g_v^{k-1})), \tag{4}$$

$$G_t^k = M(P_{\text{update}}(G_t^{k-1}, g_t^{k-1})), \tag{5}$$

$$G_m^k = M(P_{\text{update}}(G_m^{k-1}, g_m^{k-1})). \tag{6}$$

Through this process, CEGC not only transforms the raw inputs V and T into structured emotion graphs G_v^k, G_t^k, and G_m^k that explicitly encode intra- and inter-modal emotional relations, but also iteratively refines these graphs using reflection from LCR. This enables the MLLM to localize modality emotion conflicts and support progressively enhanced, interpretable emotion reasoning.

3.3 Reward-guided Conflict Feedback (RCF)

The goal of RCF is to evaluate the quality of the constructed emotion graphs (G_v^k, G_t^k, G_m^k) and to reveal unresolved modality emotion conflicts and structural deficiencies through language-based feedback. RCF contains two steps: **(1) Reward-based Graph Quality Scoring**, where a frozen human-aligned reward model provides a quality score for each graph by considering its emotional coherence together with structural validity (*e.g.*, consistency of relations and completeness under the predefined schema); and **(2) Language-level Conflict Feedback Generation**, where a frozen MLLM converts the scoring results into an evaluation message T_{eval}^k. This message highlights potential cross-modal conflicts and missing/weak links in the graphs without any parameter updates, and serves as the guidance signal for the refinement stage. By jointly scoring and verbalizing graph quality, RCF pinpoints conflicts that may reflect concealed or competing emotional evidence across modalities.

3.3.1 Step 1: Reward-Based Graph Quality Scoring. To quantify how well each emotion graph captures the modality-level emotional evidence and its conflicts, we employ a frozen human-aligned reward model R to score the video, text, and cross-modal graphs, *i.e.*, $R(G_v^k)$, $R(G_t^k)$, and $R(G_m^k)$. The score reflects overall graph quality, including emotional coherence and structural integrity, and is used in a relative manner [13] to select the most reliable graph among $\{G_v^k, G_t^k, G_m^k\}$ under the same constraints. Specifically, the top-scored graph is labeled as *Accept*, while the others are labeled as *Reject*, yielding a labeled set:

$$C^k = \sum_{i \in v,t,m} \left(G_i^k, , r_i^k \right), \tag{7}$$

where $r_i^k \in \{Accept, Reject\}$ denotes the relative evaluation outcome for each graph.

3.3.2 Step 2: Language-Level Conflict Feedback Generation. To further identify unresolved modality emotion conflicts and structural deficiencies, we apply an evaluation prompt P_{eval}, which guides the frozen MLLM to describe the strengths of accepted emotion graphs and pinpoint unresolved modality emotion conflicts or structural deficiencies in rejected ones. This produces the language-level evaluation signal $T_{\text{eval}}^{\text{k}}$:

$$T_{\text{eval}}^{\text{k}} = M(P_{\text{eval}}(C^{\text{k}})), \tag{8}$$

which provides interpretable guidance for graph refinement in the next stage.

Through RCF, emotion graphs are systematically evaluated and translated into structured language feedback. This provides a training-free signal that guides graph refinement and enhances the model's reasoning under modality emotion conflicts.

3.4 Language-driven Conflict Refinement (LCR)

The goal of LCR is to simulate reflective reasoning by transforming the language-level evaluation signal into interpretable graph editing instructions. Operating in a training-free manner, LCR leverages a frozen MLLM to analyze unresolved modality emotion conflicts and generate symbolic suggestions that guide structural refinement of the emotion graphs.

At each iteration k, LCR receives the language-level evaluation signal T_{eval}^k from RCF, which describes the strengths and weaknesses of the current emotion graphs. Using a reflection prompt P_{ref}, LCR queries the MLLM to produce language-guided graph editing signals, including operations such as removing redundant or conflicting edges, refining emotional labels, or adding missing nodes that reflect subtle emotional cues. The process is formalized as:

$$g_v^k, g_t^k, g_m^k = M(P_{\text{ref}}(T_{\text{eval}}^k)), \tag{9}$$

where g_v^k, g_t^k, and g_m^k denote the language-guided graph editing signal for the video, text, and cross-modal emotion graphs, respectively.

The generated signals are then passed to the CEGC in the next iteration to refine the emotion graphs. After $K = 2$ iterations, CERA yields refined emotion graphs that capture emotional structures with higher fidelity and reduced modality emotion conflicts. These refined graphs are finally combined with the original multimodal input (V, T) and passed to the MLLM for final prediction.

4 Experiment

4.1 Datasets

We conducted experiments on two widely used multimodal emotion datasets: MAFW [16] and CH-SIMS [32].

MAFW contained 10,045 video clips from movies, TV dramas, and social media, annotated with 11 basic and 32 compound emotion categories, along with corresponding textual descriptions. Following standard protocol, we performed five-fold cross-validation for main comparisons and used the first fold for ablations and visualizations.

CH-SIMS was a Chinese multimodal emotion dataset consisting of 2,281 video clips from diverse real-world sources. Each sample was labeled with a emotion score ranging from -1 (strongly negative) to 1 (strongly positive), with fine-grained annotations for visual, acoustic, and textual modalities.

4.2 Evaluation Metrics

Following prior work [16], we adopt standard evaluation metrics for each dataset. For the MAFW dataset, we report Unweighted Average Recall (UAR), Weighted Average Recall (WAR), and macro-averaged F1-score. UAR computes the mean recall across all emotion categories. WAR reflects overall accuracy weighted by class frequency. The macro F1-score measures the harmonic mean of precision and recall across all classes. Performance gains across these metrics reflect the robustness and generalizability of the model. For the CH-SIMS dataset, we follow prior protocols and report 2-class accuracy (Acc-2), 3-class accuracy (Acc-3), and macro-averaged F1-score to evaluate binary and ternary sentiment classification performance.

4.3 Implementation Details

Following prior work [16], we adopt standard evaluation metrics for each dataset. For MAFW and MER23, we report Unweighted Average Recall (UAR), Weighted Average Recall (WAR), and macro-averaged F1-score. For CH-SIMS, we report 2-class accuracy (Acc-2), 3-class accuracy (Acc-3), and macro-averaged F1-score. All experiments were conducted on a Linux platform using a single NVIDIA

A100 Tensor Core GPU, with implementation based on the PyTorch framework. We adopted InternVL3-8B [40] as the backbone model and utilized FsfairX-LLaMA 3-RMv0.1 [20] as the reward model. For each video, four frames were uniformly sampled and resized to 224×224 for processing. We set the number of refinement iterations to $K = 2$ based on validation results, as it offers the best balance between accuracy and stability.

4.4 Compared Methods

To evaluate the effectiveness of CERA, we compare it against a broad set of representative methods, categorized into training-based and training-free methods.

Training-based methods. On the MAFW dataset, we compare with transformer-based and fusion-enhanced models including T-ESFL [16], T-MEP [37], AMH [31], and HiCMAE-S [23], which leverage adaptive or hierarchical fusion strategies across visual and audio modalities. MAFW-DFEW-SFT [39] further fine-tunes a multimodal large language model on both the MAFW and DFEW datasets to enhance its cross-modal emotion understanding and generalization ability. On CH-SIMS, we include widely used multimodal sentiment analysis models such as TFN [35], LMF [18], MulT [24], MISA [11], MAG-BERT [22], Self-MM [34], ALMT [36], and DEVA [28], all of which are supervised models that require modality-specific training and fusion.

Training-free methods. These methods use large multimodal language models (MLLMs) in zero-shot settings without any task-specific fine-tuning. We include Qwen2-VL-7B [25], Qwen2.5-Omini [2] and Janus-Pro-1B [3], as well as the InterVL2.5 and InterVL3 series [4, 40], which represent recent strong MLLM methods. In addition, we include emotion-related MLLMs such as AffectGPT [15], EMER-SFT [39], Emotion-LLaMA [6], and MoSEAR [9], which incorporate emotional supervision or conflict-aware modeling for enhanced affective understanding. These models process video–text inputs in a unified manner, and while some incorporate affective cues or conflict-aware mechanisms, most still lack explicit and interpretable modeling of modality emotion conflicts.

4.5 Comparison with State-of-the-Art Methods

We compare the proposed CERA with a diverse set of state-of-the-art methods, categorized into training-based and training-free methods, on both MAFW and CH-SIMS datasets.

Results on MAFW. As shown in Table 1, CERA achieves the best overall performance among all training-free methods, surpassing the strong method InterVL3-8B[40] by +2.29% WAR, +2.21% UAR, and +3.23% F1. Compared to other training-free MLLMs such as Qwen2.5-Omini[2] and Janus-Pro-1B[3], CERA demonstrates substantial gains across all metrics. Although training-based models like HiCMAE-S [23] achieve competitive WAR and UAR scores, they require supervised fine-tuning and modality-specific design, whereas CERA operates in a fully training-free and generalizable manner.

Results on CH-SIMS. As shown in Table 2, CERA consistently outperforms all training-free methods, achieving 88.14% (ACC-2), 68.71% (ACC-3), and 88.13% (F1). Compared to the strong method InterVL3-8B[40], CERAimproves ACC-3 by +6.57% and ACC-2 by +1.28%. Notably, CERA also exceeds most training-based methods, including TFN[35], MulT[24], and MAG-BERT[22], and even slightly outperforms the best-performing training-based method DEVA[28] in ACC-2 and F1, highlighting its robustness in capturing nuanced emotional signals. These results validate the effectiveness of CERA in addressing modality emotion conflicts and reasoning over complex multimodal emotion cues without requiring any additional training or fine-tuning.

Table 1: Comparison results on the MAFW dataset. The best results are in bold.

Training-based Methods	WAR	UAR	F1
T-ESFL	50.29	31.00	–
T-MEP	51.15	37.17	–
AMH	48.83	32.98	–
HiCMAE-S	**54.45**	**41.66**	–
MAFW-DFEW-SFT	50.44	30.39	–
Training-free Methods	WAR	UAR	F1
Qwen2-VL-7B	43.58	36.82	47.68
Janus-Pro-1B	39.36	24.28	32.30
Qwen2.5-Omini	47.41	29.46	43.39
InternVL2.5-2B	52.76	40.66	53.75
InternVL2.5-4B	49.55	39.67	50.51
InternVL3-8B	53.54	42.28	52.47
HumanOmni-0.5B	20.18	13.52	–
EMER-SFT	38.39	28.02	–
MoSEAR	33.66	26.858	33.04
CERA (Ours)	**55.83**	**44.49**	**55.70**

Table 2: Comparison results on the CH-SIMS dataset. The best results are in bold.

Training-based Methods	ACC-2	ACC-3	F1
TFN	78.38	65.12	78.62
LMF	77.77	64.68	77.88
MulT	78.56	64.77	79.66
MISA	76.54	–	76.59
MAG-BERT	74.44	–	71.75
Self-MM	77.64	64.68	77.85
ALMT	78.56	64.98	78.94
DEVA	**79.64**	**65.42**	**80.32**
Training-free Methods	ACC-2	ACC-3	F1
Qwen2-VL-7B	76.55	48.49	75.68
Janus-Pro-1B	61.08	60.83	61.83
Qwen2.5-Omini	86.60	64.99	86.71
InternVL2.5-2B	65.98	54.05	74.44
InternVL2.5-4B	86.08	66.30	86.20
InternVL3-8B	86.86	62.14	87.00
Emotion-LLaMA	78.32	–	78.61
AffectGPT	87.20	–	87.31
MoSEAR	69.07	52.74	62.13
CERA (Ours)	**88.14**	**68.71**	**88.13**

4.6 Ablation Studies

4.6.1 The Effectiveness of Different Modules. For ablation studies, we conducted a component-wise ablation study in Table 3. We performed the ablation study on the first fold of MAFW to assess the

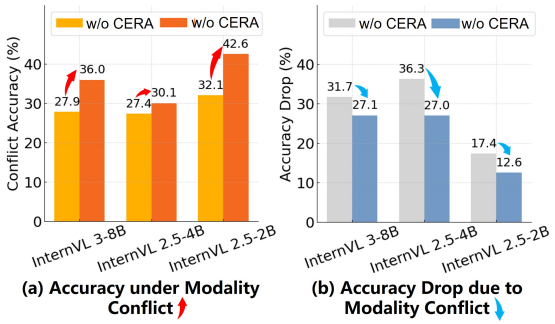

Figure 3: Evaluation of Modality Emotion Conflicts Handling in Various MLLMs. (a) Emotion recognition accuracy under modality emotion conflict scenarios. (b) The performance degradation caused by modality emotion conflicts, defined as the accuracy gap between non-conflict and conflict scenarios.

contribution of each CERA module. As shown in Table 3, adding CEGC to the baseline (InternVL3-8B) improved WAR from 47.31% to 48.09%, confirming the benefit of structured emotional graph construction. Introducing RCF further boosted WAR to 48.83% and UAR to 38.76%, demonstrating the value of language-level feedback. With all three modules, including LCR, CERA achieved the best performance (WAR: 49.97%, F1: 51.71%), validating the full perception–evaluation–reflection loop for robust reasoning.

Table 3: The effect of different modules on model performance on the first fold of MAFW database. The best results are in bold.

Baseline	CEGC	RCF	LCR	WAR	UAR	F1
✓				47.31	37.24	48.42
✓	✓			48.09	37.41	49.55
✓	✓	✓		48.83	38.76	49.82
✓	✓	✓	✓	**49.97**	**39.50**	**51.71**

4.6.2 Evaluation of Modality Emotion Conflicts Handling in Various MLLMs. To demonstrate the effectiveness of CERA in addressing modality emotion conflicts, we compare recognition performance with and without CERA under conflict scenarios across three MLLMs (InternVL3-8B [40], InternVL2.5-4B [4], and InternVL2.5-2B [4]). As shown in Figure 3, CERA consistently improves accuracy under modality emotion conflict conditions while reducing performance degradation. Figure 3(a) presents the accuracy in modality emotion conflict scenarios. CERA achieves gains of 8.1%, 2.7%, and 10.5% on InternVL3-8B, InternVL2.5-4B, and InternVL2.5-2B, respectively, showing its effectiveness in capturing and resolving cross-modal inconsistencies that MLLMs fail to handle. Figure 3(b) reports the performance degradation, defined as the accuracy gap between non-conflict and conflict scenarios. CERAreduces this degradation by 4.6% on InternVL3-8B, 9.3% on InternVL2.5-4B, and 4.8% on InternVL2.5-2B, demonstrating its robustness in preserving emotional understanding under conflicting cues. These confirm that CERA improves both accuracy and robustness by explicitly modeling and leveraging modality emotion conflicts, rather than discarding them.

4.6.3 Confusion-Matrix Evidence of Conflict Resolution. To visualize how CERA improves emotion reasoning under modality conflicts, we compare the confusion matrices of the baseline model (InternVL3-8B) and the CERA-enhanced model on the MAFW dataset. The matrices reveal class-wise misclassification patterns against ground truth. As shown in Fig. 4, all matrices are row-normalized, so diagonal entries indicate correct alignment while off-diagonals quantify class-wise confusions. Before CERA (Text × Vision), large off-diagonal mass appears between opposite-polarity categories, reflecting conflict-induced mismatches across modalities. After CERA (Vision/Text × Final), mass collapses toward the diagonal and the highlighted opposite-polarity cells shrink, yielding a higher diagonal concentration and a lower opposite-polarity confusion rate, which aligns with our quantitative gains on MAFW. These visuals align with our quantitative results, showing that CERA improves accuracy and strengthens interpretability by resolving cross-modal inconsistencies.

4.6.4 The Effectiveness of CERA across Different Baselines on the SIMS Conflict Subset. To evaluate the ability of CERA to handle modality emotion conflicts, we construct a conflict subset of CH-SIMS by selecting samples whose unimodal annotations exhibit inconsistent emotional tendencies. This subset enables a focused assessment of CERA under cross-modal contradiction scenarios. As shown in Table 4, we evaluate three representative backbones, Qwen2-VL-7B, InternVL2.5-4B, and InternVL3-8B, under zero-shot and zero-shot-CoT settings. Across all baselines, CERA consistently improves performance on the conflict subset, achieving gains of 8.96%, 7.92%, and 5.84% in ACC-2 over zero-shot for the three baselines, together with notable increases in F1. Even relative to the stronger zero shot CoT setting, CERA further boosts ACC-3 and F1, for example plus 7.34% in ACC-3 on InternVL3-8B, confirming its superior conflict resolution capability. While CERA introduces a modest runtime increase from 1.1x to 1.4x compared with zero-shot-CoT on the same hardware, it maintains comparable GPU memory usage. These results indicate that the proposed perception–evaluation–reflection mechanism enhances robustness to modality conflicts without any additional training and with minimal computational overhead.

Table 4: Comparison results of CERA across different baselines on the SIMS conflict subset. The best results are in bold.

Methods	Memory	ACC-2	ACC-3	F1
Qwen2-VL-7B				
zero-shot	7340MiB	55.97	42.37	60.73
zero-shot-CoT	7370MiB	62.89	42.94	61.90
+CERA	7312MiB	**64.93**	**50.28**	**64.90**
InterVL2.5-4B				
zero-shot	10516MiB	67.91	47.46	67.60
zero-shot-CoT	14784MiB	69.40	48.59	69.60
+CERA	13394MiB	**73.13**	**49.72**	**72.19**
InterVL3-8B				
zero-shot	18582MiB	68.79	45.76	69.04
zero-shot-CoT	19502MiB	70.90	46.89	72.72
+CERA	19618MiB	**74.63**	**51.41**	**74.49**

4.6.5 The Impact of Absent Modalities. To investigate the importance of each modality, we conduct an ablation study by selectively

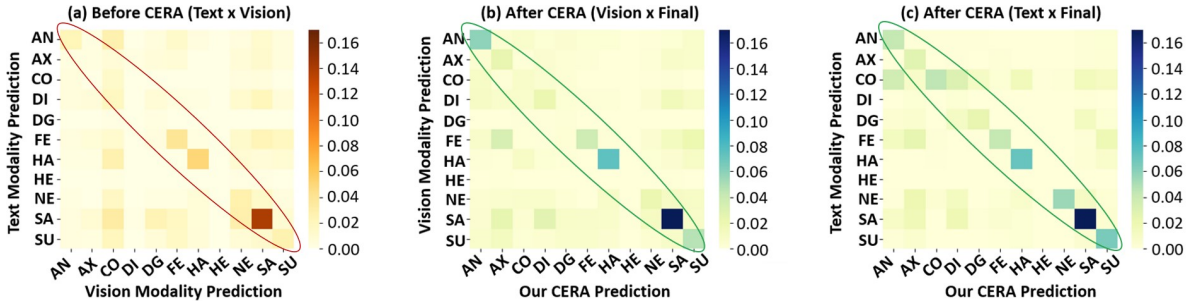

Figure 4: Confusion-matrix evidence of conflict resolution on the first fold of MAFW. (a) visualizes disagreements between the text-modality prediction (rows) and the vision-modality prediction (columns); deep off-diagonal blocks between emotionally opposite categories indicate conflict-induced misclassifications. (b) and (c) compare each unimodal prediction with CERA's prediction. In both (b) and (c), mass concentrates on the diagonal while opposite-polarity confusions (boxed cells) are markedly reduced. This diagonal consolidation, together with lower opposite-polarity confusion rates, demonstrates that CERA resolves cross-modal inconsistencies and stabilizes class-wise decisions. Note: AN (anger), AX (anxiety), CO (contempt), DI (disappointment), DG (disgust), FE (fear), HA (happiness), HE (helplessness), NE (neutral), SA (sadness), SU (surprise).

removing the video (V) or text (T) modality while keeping the rest of the CERA pipeline unchanged. As shown in Table 5, removing either modality leads to a significant drop in performance across all three metrics, highlighting the complementary nature of video and text in emotion understanding. Specifically, removing the text modality (*w/o T*) causes the largest performance degradation, with WAR dropping from 49.97% to 43.39%, suggesting that text often carries decisive emotional signals. On the other hand, removing the video modality (*w/o V*) also leads to noticeable drops, especially in F1, indicating that visual cues contribute to the model's balanced recognition across emotion classes. These results underscore the necessity of both modalities for robust and fine-grained multimodal emotion reasoning.

Table 5: The effect of different modalities on the first fold of MAFW database. The best results are in bold.

Method	WAR	UAR	F1
w/o V	48.18	38.46	43.05
w/o T	43.39	34.54	42.81
CERA(Ours)	**49.97**	**39.50**	**51.71**

4.6.6 Comparison with Existing Conflict-Aware Methods on the SIMS Conflict Subset. To further examine CERA's effectiveness in resolving modality emotion conflicts, we compare it with two representative conflict-aware approaches, Self-MM and MoSEAR, as shown in Table 6. Both methods are specifically designed to address modality inconsistency: Self-MM [34] learns modality-specific and invariant representations through self-supervised multi-task learning, while MoSEAR [9] introduces conflict-aware regularization to mitigate emotional inconsistencies during multimodal fusion. Despite their targeted designs, both methods still rely on task-specific training. In contrast, CERA achieves the best results across all metrics under a fully training-free setting, reaching 74.63% ACC-2, 51.41% ACC-3, and 74.49% F1. Compared with Self-MM, CERA improves ACC-3 and F1 by +3.39% and +5.04%, respectively, demonstrating that its reflective reasoning loop more effectively captures and resolves cross-modal contradictions. These results validate that CERA provides a general and interpretable alternative to existing conflict-aware frameworks, achieving superior robustness to modality emotion conflicts without requiring any model training or parameter optimization.

Table 6: The comparison with other conflict-aware methods on the SIMS conflict subset. The best results are in bold.

Method	ACC-2	ACC-3	F1
training-based			
Self-MM	69.49	48.02	69.45
training-free			
MoSEAR	55.22	38.42	44.99
CERA(Ours)	**74.63**	**51.41**	**74.49**

5 Conclusion

In this paper, we present **Conflict-Explicit Reflective Agent (CERA)**, a novel training-free, conflict-aware, and language-driven agentic framework for multimodal emotion reasoning. Unlike existing methods that treat modality emotion conflicts as noise, CERA explicitly models and resolves such conflicts as meaningful emotional signals, enhancing both recognition accuracy and interpretability. CERA operates through a structured perception–evaluation–reflection loop consisting of three stages: (1) Conflict-perceptive Emotion Graph Construction (CEGC) constructs fine-grained emotion graphs to reveal and iteratively refine conflicts; (2) Reward-guided Conflict Feedback (RCF) uses a reward model to evaluate graphs and generate natural-language feedback on unresolved conflicts; and (3) Language-driven Conflict Refinement (LCR) transforms the feedback into graph-level updates without any parameter tuning, allowing CERA to progressively refine its emotional reasoning. Experiments on MAFW and CH-SIMS show CERA outperforms state-of-the-art methods in both accuracy and interpretability. Future work will extend CERA to audio and explore adaptive, human-in-the-loop refinement mechanisms.

Acknowledgments

This work was supported by the National Natural Science Foundation of China grant (62076227), Natural Science Foundation of Hubei Province grant (2023AFB572) and Hubei Key Laboratory of Intelligent Geo-Information Processing (KLIGIP-2022-B10).

References

[1] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Iryna Gurevych and Yusuke Miyao (Eds.). Association for Computational Linguistics, Melbourne, Australia, 2236–2246. doi:10.18653/v1/P18-1208

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-VL Technical Report. arXiv:2502.13923 [cs.CV] https://arxiv.org/abs/2502.13923

[3] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. 2025. Janus-Pro: Unified Multimodal Understanding and Generation with Data and Model Scaling. arXiv:2501.17811 [cs.AI] https://arxiv.org/abs/2501.17811

[4] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2025. Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling. arXiv:2412.05271 [cs.CV] https://arxiv.org/abs/2412.05271

[5] Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Jingdong Sun, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander G. Hauptmann. 2024. Emotion-LLaMA: Multimodal Emotion Recognition and Reasoning with Instruction Tuning. In *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (Eds.), Vol. 37. Curran Associates, Inc., 110805–110853. https://proceedings.neurips.cc/paper_files/paper/2024/file/c7f43ada17acc234f568dc66da527418-Paper-Conference.pdf

[6] Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander Hauptmann. 2024. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. *Advances in Neural Information Processing Systems* 37 (2024), 110805–110853.

[7] Yumeng Fu, Junjie Wu, Zhongjie Wang, Meishan Zhang, Lili Shan, Yulin Wu, and Bingquan Li. 2025. LaERC-S: Improving LLM-based Emotion Recognition in Conversation with Speaker Characteristics. arXiv:2403.07260 [cs.CL] https://arxiv.org/abs/2403.07260

[8] Markus Hafner, Maria Katsantoni, Tino Köster, James Marks, Joyita Mukherjee, Dorothee Staiger, Jernej Ule, and Mihaela Zavolan. 2021. CLIP and complementary methods. *Nature Reviews Methods Primers* 1, 1 (2021), 20.

[9] Zhiyuan Han, Beier Zhu, Yanlong Xu, Peipei Song, and Xun Yang. 2025. Benchmarking and bridging emotion conflicts for multimodal emotion reasoning. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 5528–5537.

[10] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis. arXiv:2005.03545 [cs.CL] https://arxiv.org/abs/2005.03545

[11] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis. In *Proceedings of the 28th ACM International Conference on Multimedia* (Seattle, WA, USA) *(MM '20)*. Association for Computing Machinery, New York, NY, USA, 1122–1131. doi:10.1145/3394171.3413678

[12] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. 2021. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. arXiv:2107.07651 [cs.CV] https://arxiv.org/abs/2107.07651

[13] Yafu Li, Xuyang Hu, Xiaoye Qu, Linjie Li, and Yu Cheng. 2025. Test-Time Preference Optimization: On-the-Fly Alignment via Iterative Textual Feedback. *arXiv e-prints* (2025), arXiv–2501.

[14] Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, Yong Ren, Zebang Cheng, Bin Liu, Rui Liu, Xiaojiang Peng, et al. 2025. Affectgpt: A new dataset, model, and benchmark for emotion understanding with multimodal large language models. *arXiv preprint arXiv:2501.16566* (2025).

[15] Zheng Lian, Haiyang Sun, Licai Sun, Jiangyan Yi, Bin Liu, and Jianhua Tao. 2024. AffectGPT: Dataset and framework for explainable multimodal emotion recognition. *arXiv preprint arXiv:2407.07653* (2024).

[16] Yuanyuan Liu, Wei Dai, Chuanxu Feng, Wenbin Wang, Guanghao Yin, Jiabei Zeng, and Shiguang Shan. 2022. MAFW: A Large-scale, Multi-modal, Compound Affective Database for Dynamic Facial Expression Recognition in the Wild. In *Proceedings of the 30th ACM International Conference on Multimedia* (Lisboa, Portugal) *(MM '22)*. Association for Computing Machinery, New York, NY, USA, 24–32. doi:10.1145/3503161.3548190

[17] Yuanyuan Liu, Yuxuan Huang, Shuyang Liu, Yibing Zhan, Zijing Chen, and Zhe Chen. 2024. Open-set video-based facial expression recognition with human expression-sensitive prompting. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 5722–5731.

[18] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2018. Efficient Low-rank Multimodal Fusion with Modality-Specific Factors. arXiv:1806.00064 [cs.AI] https://arxiv.org/abs/1806.00064

[19] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, and S. Poria. 2018. Multimodal sentiment analysis using hierarchical fusion with context modeling. *Knowledge-Based Systems* 161 (2018), 124–133. doi:10.1016/j.knosys.2018.07.041

[20] Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems* 37 (2024), 124198–124235.

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 [cs.CV] https://arxiv.org/abs/2103.00020

[22] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating Multimodal Information in Large Pretrained Transformers. *Proceedings of the conference. Association for Computational Linguistics. Meeting* 2020 (July 2020), 2359—2369. doi:10.18653/v1/2020.acl-main.214

[23] Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. 2024. HiCMAE: Hierarchical Contrastive Masked Autoencoder for self-supervised Audio-Visual Emotion Recognition. *Information Fusion* 108 (2024), 102382. doi:10.1016/j.inffus.2024.102382

[24] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal Transformer for Unaligned Multimodal Language Sequences. *Proceedings of the conference. Association for Computational Linguistics. Meeting* 2019 (July 2019), 6558—6569. doi:10.18653/v1/p19-1656

[25] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. arXiv:2409.12191 [cs.CV] https://arxiv.org/abs/2409.12191

[26] Wenbin Wang, Liang Ding, Li Shen, Yong Luo, Han Hu, and Dacheng Tao. 2024. WisdoM: Improving Multimodal Sentiment Analysis by Fusing Contextual World Knowledge. In *Proceedings of the 32nd ACM International Conference on Multimedia* (Melbourne VIC, Australia) *(MM '24)*. Association for Computing Machinery, New York, NY, USA, 2282–2291. doi:10.1145/3664647.3681403

[27] Xincheng Wang, Liejun Wang, Yinfeng Yu, and Xinxin Jiao. 2025. Modality-Invariant Bidirectional Temporal Representation Distillation Network for Missing Multimodal Sentiment Analysis. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[28] Sheng Wu, Dongxiao He, Xiaobao Wang, Longbiao Wang, and Jianwu Dang. 2025. Enriching Multimodal Sentiment Analysis Through Textual Emotional Descriptions of Visual-Audio Content. *Proceedings of the AAAI Conference on Artificial Intelligence* 39, 2 (Apr. 2025), 1601–1609. doi:10.1609/aaai.v39i2.32152

[29] Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025. Qwen2.5-Omni Technical Report. arXiv:2503.20215 [cs.CL] https://arxiv.org/abs/2503.20215

[30] Qize Yang, Detao Bai, Yi-Xing Peng, and Xihan Wei. 2025. Omni-emotion: Extending video mllm with detailed face and audio modeling for multimodal emotion analysis. *arXiv preprint arXiv:2501.09502* (2025).

[31] Seunghyun Yoon, Subhadeep Dey, Hwanhee Lee, and Kyomin Jung. 2020. Attentive Modality Hopping Mechanism for Speech Emotion Recognition. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 3362–3366. doi:10.1109/ICASSP40776.2020.9054229

[32] Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. 2020. CH-SIMS: A Chinese Multimodal Sentiment Analysis Dataset with Fine-grained Annotation of Modality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 3718–3727. doi:10.18653/v1/2020.acl-main.343

[33] Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 10790–10797.

[34] Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. Learning Modality-Specific Representations with Self-Supervised Multi-Task Learning for Multimodal Sentiment Analysis. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 12 (May 2021), 10790–10797. doi:10.1609/aaai.v35i12.17289

[35] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor Fusion Network for Multimodal Sentiment Analysis.

arXiv:1707.07250 [cs.CL] https://arxiv.org/abs/1707.07250

[36] Haoyu Zhang, Yu Wang, Guanghao Yin, Kejun Liu, Yuanyuan Liu, and Tianshu Yu. 2023. Learning Language-guided Adaptive Hyper-modality Representation for Multimodal Sentiment Analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. doi:10.18653/v1/2023.emnlp-main.49

[37] Xiaoqin Zhang, Min Li, Sheng Lin, Hang Xu, and Guobao Xiao. 2024. Transformer-Based Multimodal Emotional Perception for Dynamic Facial Expression Recognition in the Wild. *IEEE Transactions on Circuits and Systems for Video Technology* 34, 5 (2024), 3192–3203. doi:10.1109/TCSVT.2023.3312858

[38] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923* (2023).

[39] Jiaxing Zhao, Xihan Wei, and Liefeng Bo. 2025. R1-omni: Explainable omni-multimodal emotion recognition with reinforcement learning. *arXiv preprint arXiv:2503.05379* (2025).

[40] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2025. InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. arXiv:2504.10479 [cs.CV] https://arxiv.org/abs/2504.10479