Multimodal Emotion Recognition with Large Language Models

Hongrui Zhang[1,2], **Daiqing Wu**[1], **Yangyang Li**[3], **Kuien Liu**[3], **Yuhui Wang**[3], **Yu Zhou**[4] and **Sicheng Zhao**[1*]

[1]Department of Psychological and Cognitive Sciences, Tsinghua University, Beijing, China
[2]School of Computer Science and Technology, Harbin Institute of Technology, Weihai, China
[3]Academy of Cyber, Beijing, China
[4]College of Computer Science, Nankai University, Tianjin, China
smilingweeping@gmail.com, schzhao@tsinghua.edu.cn

Abstract

Multimodal Emotion Recognition (MER) focuses on identifying and interpreting emotions from modality-compound inputs. Closely mirroring human cognitive processes in real-world environments, MER has drawn substantial attention from both academia and industry. Recently, a paradigm shift has been unveiled in MER, from leveraging small-scale, task-specific models to Large Language Models (LLMs). We refer to the latter as the MER-with-LLMs paradigm, which offers unprecedented generality, spurring numerous empirical attempts, even alongside speculation about LLMs' potential to achieve general emotional intelligence. However, with these new opportunities come new challenges, including the scarcity of emotionally annotated data, the affective gap both within and across modalities, and the opacity of affective interpretation. To systematically review existing research and guide future exploration, this paper categorizes prior works according to their focus on addressing these challenges into three directions: Affective Data Augmentation, Multimodal Affective Representation, and Multimodal Affective Reasoning. By thoroughly tracing the development, emerging trends, and remaining issues within each direction, this paper aims to provide a clear academic map of the MER-with-LLMs paradigm and foster its structured advancement.

1 Background and Challenges

Emotion is an integral component of human daily experiences, significantly influencing individuals' communication, decision-making, and behavior. In real-world settings, emotions are expressed and perceived through multiple modalities, including language, speech, facial expressions, and gestures [Zhao *et al.*, 2023; Wang *et al.*, 2026b]. As a result, Multimodal Emotion Recognition (MER) has emerged as a critical research area, focusing on enabling models to comprehend emotions from modality-compound inputs as humans do. To achieve this, traditional research primarily relies on small-scale, task-specific models [Yang *et al.*, 2025b; Li *et al.*, 2025c]. Despite achieving satisfactory performance, they suffer from strong dependencies on predefined input domains and output spaces, which make them less effective in practical applications that require dynamism and flexibility.

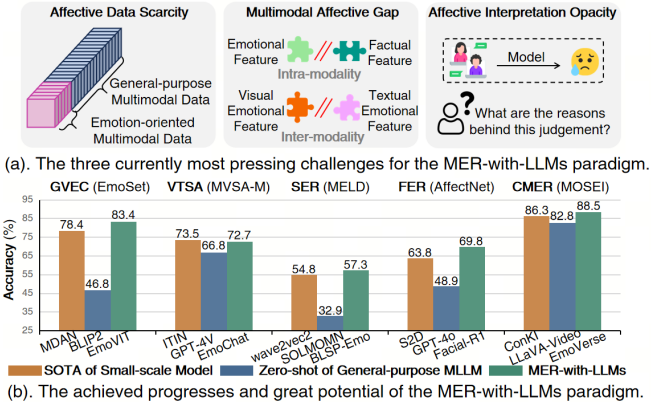

Figure 1: The MER-with-LLMs paradigm faces three major challenges, along with significant potential and opportunities.

The emerging Large Language Models (LLMs) present a promising avenue for addressing these limitations [Shou *et al.*, 2025]. Large-scale pre-training equips LLMs with strong instruction-following capabilities, which are inherited by Multimodal LLMs (MLLMs), giving rise to a new **MER-with-LLMs** paradigm. By formulating MER as an LLM-centric autoregressive process, this paradigm enables unified handling of inputs spanning multiple domains and modalities, as well as producing diverse instruction-conditioned outputs. It even deepens the MER task itself, extending classification toward explanation [Lian *et al.*, 2024a]. Collectively, these advantages have driven a paradigm shift, reflected by a rapidly growing body of recent work on MER-with-LLMs.

Alongside this trend, new challenges have also arrived. A line of studies [Lian *et al.*, 2024b; Wu *et al.*, 2025a] has reached a consensus that, under zero-shot inference, MLLMs often fail to achieve proficiency on MER comparable to that observed in other multimodal tasks. This deficiency is inherently rooted in the complexity of MER, which entails capturing high-level affective cues from multimodal inputs, modeling interactions among heterogeneous modalities, and inte-

*Corresponding author.

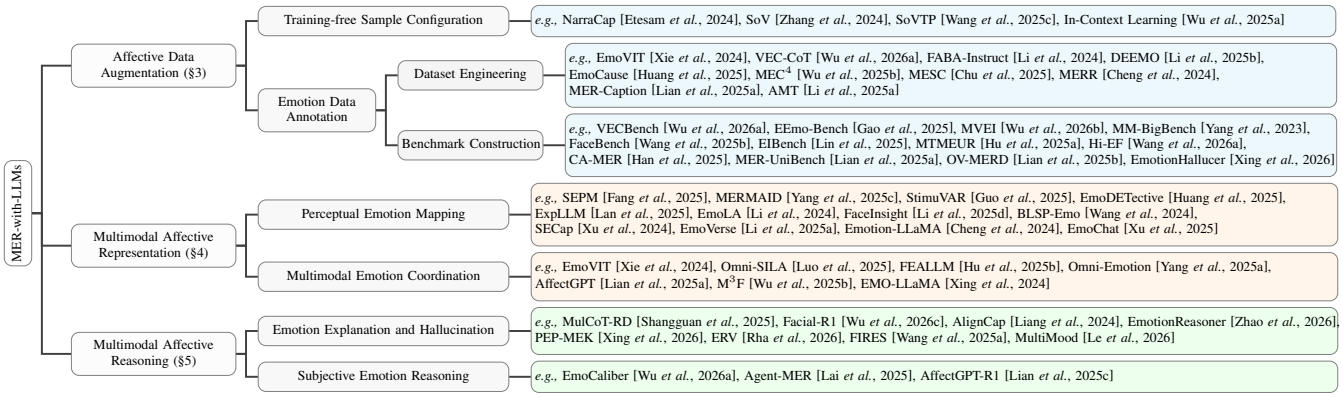

Figure 2: Taxonomy of the MER-with-LLMs paradigm. We categorize existing works into three branches based on their primary challenges, with representative works listed on the leaves.

grating them to derive emotional conclusions. Following a progressive order, we categorize the currently most pressing challenges into three branches, as illustrated in Fig. 1(a):

Affective Data Scarcity. Data constitute the foundation of modern models. The underperformance of general MLLMs on MER underscores the importance of emotional data, making targeted optimization on such data a direct and intuitive remedy. However, the subjectivity of emotion perception often necessitates collaboration from multiple annotators for reliable annotation, substantially increasing labeling costs and constraining the scale of high-quality datasets. Moreover, datasets collected in different contexts are typically annotated under distinct and often incompatible emotion theories, further aggravating dataset fragmentation. These limitations compel existing approaches to either maximize the utilization of available data or construct new datasets.

Multimodal Affective Gap. The "affective gap" generally refers to the intra-modality misalignment between emotional and factual features [Zhao *et al.*, 2021b], which challenges MLLMs in capturing meaningful affective cues. We extend this concept to cover an inter-modality setting, where it characterizes the heterogeneity and semantic discrepancy of emotional features across different modalities. Heterogeneity reflects inconsistencies in information density and expression patterns; for instance, text tends to be more compact and abstract, whereas images are more dispersed and concrete. Semantic discrepancy refers to the fact that each modality provides unique and sometimes conflicting cues. Together, these factors complicate inter-modality emotional synergy modeling. Overall, these gaps motivate the exploration of dedicated affective cues extraction and interaction modeling.

Affective Interpretation Opacity. Inherited from the habits of small-scale models, mainstream MER methods primarily focus on emotion recognition rather than explanation. This emphasis results in a lack of decision transparency, which not only hinders iterative model improvement but also raises concerns about reliability in high-stakes and cross-domain applications, including healthcare, education, and finance. Benefiting from the autoregressive generation process of LLMs, the MER-with-LLMs paradigm naturally supports step-by-step reasoning or explicit natural language explanations for

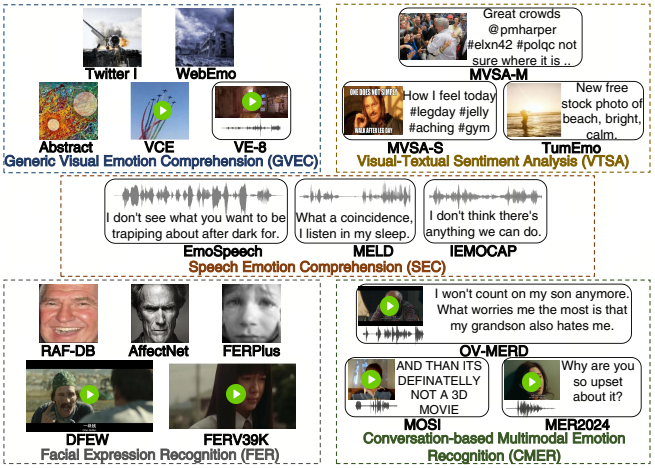

Figure 3: Visualization of samples from representative datasets of the five mainstream MER sub-tasks.

emotional decisions, thereby giving rise to a new task: explainable MER. In this context, existing approaches aim to enhance model transparency and reliability, for example, by reducing hallucinations, eliciting confidence verbalization, and enabling open-vocabulary MER.

To address these challenges, various optimization strategies with divergent emphases have been proposed, significantly advancing this paradigm, as illustrated in Fig. 1(b). However, their optimization pathways remain intricate and diversified, and a systematic organization is still lacking. To promote a more structured development of this emerging paradigm, this paper provides a comprehensive review of the existing challenges, a taxonomy of current methods, and an in-depth discussion of recent progress and future directions.

2 Problem Formulation and Taxonomy

Compared with traditional MER [Zhao *et al.*, 2021a], we extend the MER-with-LLMs paradigm to a broader setting. Under this paradigm, LLMs serve as a text-centered orchestrator that processes affective information from multimodal inputs containing at least one non-textual modality.

	Method	Modality	Sub-task	#Sample	#Labels	Annotators	Key Focus
Training Dataset	EmoVIT [Xie *et al.*, 2024]	I,T	GVEC	51,200	-	Model	Emotional Instruction
	VEC-CoT [Wu *et al.*, 2026a]	I,T	GVEC	143,446	-	Model	Structured Affective Reasoning
	FABA-Instruct [Li *et al.*, 2024]	I,T	FER	19,877	7	Model-led+Human-assisted	Face Analysis
	DEEMO-MER [Li *et al.*, 2025b]	V,A,T	CMER	2,060	2	Model-led+Human-assisted	Identity Privacy
	DEEMO-NFBL [Li *et al.*, 2025b]	V,A,T	CMER	24,722	37	Human	Identity Privacy
	EmoCause [Huang *et al.*, 2025]	V,A,T	CMER	14,000	8	Model-led+Human-assisted	Emotional Cause Understanding
	MEC^4 [Wu *et al.*, 2025b]	V,A,T	CMER	2,124	-	Human	Emotional Cause Understanding
	MESC [Chu *et al.*, 2025]	V,A,T	CMER	28,762	7	Human-led+Model-assisted	Emotional Support
	MERR-Coarse [Cheng *et al.*, 2024]	V,A,T	CMER	28,618	-	Model	Multimodal Emotion Instruction
	MERR-Fine [Cheng *et al.*, 2024]	V,A,T	CMER	4,487	-	Human-led+Model-assisted	Multimodal Emotion Instruction
	MER-Caption+ [Lian *et al.*, 2025a]	V,A,T	CMER	31,327	-	Model-led+Human-assisted	Multimodal Emotion Instruction
	MER-Caption [Lian *et al.*, 2025a]	V,A,T	CMER	115,595	-	Model-led+Human-assisted	Multimodal Emotion Instruction
	AMT [Li *et al.*, 2025a]	I,V,T	FER,CMER	32,940	10	Human-led+Model-assisted	Affective Multitask
Evaluation Benchmark	VECBench [Wu *et al.*, 2026a]	I	GVEC	8,235	-	Human	Systematic Organization
	EEmo-Bench [Gao *et al.*, 2025]	I,T	GVEC	6,733	7	Human	Image-evoked Emotions
	MVEI [Wu *et al.*, 2026b]	I,T	GVEC	3,086	OV	Model-led+Human-assisted	Emotion Statement Judgment
	MM-BigBench [Yang *et al.*, 2023]	I,T	VTSA	31,902	-	Human	Systematic Organization
	FaceBench [Wang *et al.*, 2025b]	I	FER	73,760	7	Human-led+Model-assisted	Face Analysis
	EIBench [Lin *et al.*, 2025]	I,T	FER	1,665	4	Model-led+Human-assisted	Emotion Interpretation
	MTMEUR [Hu *et al.*, 2025a]	V,T	CMER	1,451	7	Model-led+Human-assisted	Multi-turn Understanding
	Hi-EF [Wang *et al.*, 2026a]	V,A,T	CMER	3,069	7	Human	Emotion Forecasting
	CA-MER [Han *et al.*, 2025]	V,A,T	CMER	1,500	9	Model-led+Human-assisted	Emotion Conflicts
	MER-UniBench [Lian *et al.*, 2025a]	V,A,T	CMER	12,799	-	Human	Systematic Organization
	OV-MERD [Lian *et al.*, 2025b]	V,A,T	CMER	332	OV	Human-led+Model-assisted	Open-vocabulary
	EmotionHallucer [Xing *et al.*, 2026]	I,V,A,T	CMER	2,742	-	Human-led+Model-assisted	Emotion Hallucinations

Table 1: Emotion Data Annotation methods with their most critical and distinguishing attributes. We consider a dataset comprises modality "T" if the text carries information beyond basic task description. "-" indicates descriptive/compound dataset. "OV" denotes open-vocabulary.

Based on this definition, we first provide a symbolic formulation before taxonomizing existing methods. In general, a multimodal sample X can consist of various modalities, such as I (image), V (video), A (audio), T (text), or even physiological signals like electroencephalogram (EEG). Although the latter holds critical application value [Cai *et al.*, 2022], this paper primarily focuses on I, V, A, and T modalities, considering their widespread availability. Together with a text instruction Q that specifies the task requirements, the sample is fed into an MLLM π_θ. To process these inputs, the MLLM typically employs modality-specific encoders and adapters to transform I, V, and A into multimodal embeddings, which are then combined with the tokenized T and Q to generate a textual response $R = [r_1, r_2, \cdots, r_n]$. Each token is generated by maximizing the conditional likelihood given the multimodal inputs and the preceding responses:

$$r_i = \arg\max_{r \in \mathcal{V}} \pi_\theta(r|X, Q, R_{<i}), \tag{1}$$

Here, $\mathcal{V}$ denotes the vocabulary of the MLLM. Depending on downstream requirements, the desired responses can take the form of emotion categories or tailored explanations. Separated by input domains and modalities, we divide MER into five mainstream sub-tasks, with samples visualized in Fig. 3:

- *Generic Visual Emotion Comprehension (GVEC)* aims to predict viewer-side emotional responses elicited by visual stimuli. $X = \{I\}/\{V\}/\{V, A\}$ consists of images or videos from arbitrary domains, such as artworks, landscapes, movies, or human-shared content.
- *Visual-Textual Sentiment Analysis (VTSA)* aims to identify the emotion embedded in image–text pairs. $X = \{I, T\}$ is typically image–text posts collected from social media platforms such as Twitter and Tumblr.

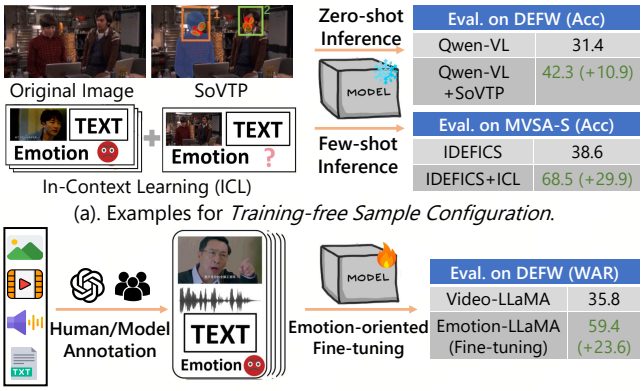

Figure 4: Illustration of Affective Data Augmentation methods.

- *Speech Emotion Comprehension (SEC)* aims to describe emotions conveyed by speeches, commonly through classification or captioning. $X = \{A\}$ is a speech segment extracted from TV shows or purposely recorded.
- *Facial Expression Recognition (FER)* aims to infer human emotional states from facial expressions. $X = \{I\}/\{V\}$ focuses on facial cues, typically in the form of a face-centered static image or dynamic video.
- *Conversation-based Multimodal Emotion Recognition (CMER)* extends FER to a conversational setting, aiming to analyze a speaker's emotional state from audiovisual dialogues. $X = \{V, A, T\}/\{V, T\}$ is typically sourced from movies or TV shows.

On top of the above formulation, we present the taxonomy of MER-with-LLMs in Fig. 2. We categorize existing meth-

Method	Modality	Sub-task	Technique	Encoder Adaptation		Key Design
				Image/Video	Audio	
SEPM [Fang *et al.*, 2025]	I	GVEC	Zero-Shot	Sharpened	-	Coarse to Fine Inference
MERMAID [Yang *et al.*, 2025c]	I	GVEC	Zero-Shot	Augmented	-	Self-reflective Agent
StimuVAR [Guo *et al.*, 2025]	V	GVEC	SFT	Sharpened	-	Recognition before Reasoning
EmoDETective [Huang *et al.*, 2025]	V,A	GVEC	SFT+RL	Basic	Basic	Fact before Emotion
ExpLLM [Lan *et al.*, 2025]	I	FER	SFT	Basic	-	Multi-turn Conversation
EmoLA [Li *et al.*, 2024]	I	FER	SFT	Augmented	-	Dual-perspective Encoding
FaceInsight [Li *et al.*, 2025d]	I	FER	SFT	Augmented	-	Correlation&Logical Constraint
BLSP-Emo [Wang *et al.*, 2024]	A	SEC	SFT	-	Basic	Unimodal before Multimodal
SECap [Xu *et al.*, 2024]	A	SEC	SFT	-	Sharpened	Q-former-based Compression
EmoVerse [Li *et al.*, 2025a]	V,T	CMER	SFT	Basic	-	Multi-task Learning
Emotion-LLaMA [Cheng *et al.*, 2024]	V,A,T	FER, CMER	SFT	Augmented	Basic	Multi-view Encoding
EmoChat [Xu *et al.*, 2025]	I,V,A,T	GVEC, VTSA, FER, CMER	SFT	Augmented	Basic	Feature Alignment

Table 2: Perceptual Emotion Mapping methods with their key attributes. "Encoder Adaptation" characterizes how each method adapts the encoding process beyond parameter updating. "Basic" indicates no adaptation. SFT: Supervised Fine-tuning. RL: Reinforcement Learning.

ods into three branches according to the challenges they focus on and provide a high-level overview below:

Affective Data Augmentation. Existing methods for handling Affective Data Scarcity are divided into two branches. *Training-free Sample Configuration* aims to maximize the utilization of available data, typically by augmenting the input context while freezing MLLM parameters. *Emotion Data Annotation* focuses on augmenting the volume of data, either by engineering training data for fine-tuning or by constructing benchmarks that enable more emotion-aware evaluation.

Multimodal Affective Representation. To handle Multimodal Affective Gap, a line of research focuses on feeding multimodal embeddings with richer affective representations into LLMs. According to their primary emphasis, we further divide these methods into two branches: *Perceptual Emotion Mapping*, which refines the perception of intra-modality affective cues by optimizing the encoding process, and *Multimodal Emotion Coordination*, which enhances the modeling of inter-modality coordination through comprehensive fusion of affective cues from multiple sources.

Multimodal Affective Reasoning. To address Affective Interpretation Opacity, some studies encourage MLLMs to generate step-by-step reasoning before producing final emotion predictions, thereby enhancing explainability. Pioneering works primarily focus on the accuracy and reliability of such reasoning, which we categorize as *Emotion Explanation and Hallucination*. More recent studies further explore accommodating the inherent subjectivity of emotion perception, which we classify as *Subjective Emotion Reasoning*.

Below, we elaborate on each branch. Methods that fall into overlapping regions are classified by their primary focus.

3 Affective Data Augmentation

Training-free Sample Configuration. To handle Affective Data Scarcity, this line of work emphasizes maximizing the utility of existing samples. This is achieved by carefully configuring the samples provided as input to MLLMs, a process that introduces human-prior-inspired biases or demonstrations. Such strategies are cost-efficient and are particularly useful when additional data is unavailable.

Under zero-shot settings, the biases are injected through extracting key image features, allowing MLLMs to focus on regions that convey the most salient emotional cues. For example, initial explorations [Etesam *et al.*, 2024] utilize CLIP [Radford *et al.*, 2021] to extract emotional information from bounding boxes around individuals in images, generating captions that reflect emotional content. While bounding boxes alone provide useful spatial localization, they often lack the granularity needed for nuanced MER. To overcome this limitation, several approaches [Zhang *et al.*, 2024; Wang *et al.*, 2025c] incorporate multi-level visual cues, including spatial localization, facial geometry, and action units, enabling a more comprehensive analysis of human emotional information. Inspired by the success of In-Context Learning (ICL), Wu *et al.* [2025a] unleash the emotion perception of MLLMs in a few-shot setting. By optimizing the retrieval, presentation, and distribution of demonstrations, this method allows MLLMs to generalize effectively from minimal labeled data, further improving emotion inference in multimodal contexts. We illustrate them in Fig. 4(a).

Emotion Data Annotation. When additional data is permitted, annotating the desired emotion-oriented datasets serves as a more straightforward approach. In this line of work, we separately introduce the training datasets and evaluation benchmarks, both of which are summarized in Tab. 1.

Dataset Engineering. This branch enhances the intrinsic capabilities of MLLMs by constructing training datasets, as illustrated in Fig. 4(b). For the GVEC task, EmoVIT [Xie *et al.*, 2024] leverages advanced proprietary models to annotate the first emotion-centric visual instruction tuning dataset, and VEC-CoT [Wu *et al.*, 2026a] subsequently contributes a larger-scale structured affective reasoning dataset. For the FER task, FABA-Instruct [Li *et al.*, 2024] considers both action units and emotions, introducing the first dataset of its kind. AMT [Li *et al.*, 2025a] further extends to a multi-task scenario. The CMER task has received the most attention, likely due to its highly complex input modalities, with different methods exploring complementary scenarios. Specifically, DEEMO [Li *et al.*, 2025b] is designed to preserve privacy; EmoCause [Huang *et al.*, 2025] and MEC^4 [Wu *et al.*, 2025b] focus on understanding emotion cause under multi-attribute and multilingual settings; MESC [Chu *et al.*, 2025] targets therapeutic and counseling scenarios; and

Method	Modality	Sub-task	Technique	Encoder	Abbreviated Fusion Mechanism
EmoVIT [Xie *et al.*, 2024]	I	GVEC	SFT	I	Q-former(q=[Learned,T],k&v=I)
Omni-SILA [Luo *et al.*, 2025]	V,T	GVEC	SFT	V,V_f,V_h,V_o	Mixture-of-Experts(V,V_f,V_h,V_o)
FEALLM [Hu *et al.*, 2025b]	I	FER	SFT	I,I_f	Cross-attention(q=I,k&v=I_f)
Omni-Emotion [Yang *et al.*, 2025a]	V,A,T	FER, CMER	SFT	V,V_f,A	[V,V_f]
AffectGPT [Lian *et al.*, 2025a]	V,A,T	CMER	SFT	V,A	Self-attention([V,A])
M^3F [Wu *et al.*, 2025b]	V,A,T	CMER	SFT	V,A	Q-former(q=Learned,k&v=V)+Q-former(q=Learned,k&v=A)
EMO-LLaMA [Xing *et al.*, 2024]	I,V,T	FER, CMER	SFT	V,V_f	Q-former(q=[Learned,T],k&v=[V,V_f])+Self-attention([V,V_f])

Table 3: Multimodal Emotion Coordination methods with their key attributes. "Encoder" lists the adopted encoders, where I_f represents image facial encoder, V_f, V_h, and V_o denote video facial, human, object encoders, respectively. Operation [·] represents concatenation.

MERR [Cheng *et al.*, 2024] and MER-Caption [Lian *et al.*, 2025a] address regular CMER tasks, characterized by their high quality and large scale. These datasets provide essential prerequisites for subsequent emotion-oriented optimization.

Benchmark Construction. Beyond training, evaluation also plays an indispensable role in the development of MER-with-LLMs, whose evolution exhibits unexpected consistency across MER sub-tasks. Pioneering efforts focus on systematically integrating existing datasets, such as VECBench [Wu *et al.*, 2026a], MM-BigBench [Yang *et al.*, 2023], and MER-UniBench [Lian *et al.*, 2025a]. Subsequent works aim to shed light on deeper insights. For example, EEmo-Bench [Gao *et al.*, 2025] unifies evaluation of perception, description, ranking, and assessment; FaceBench [Wang *et al.*, 2025b] encompasses perception for fine-grained facial details; EIBench [Lin *et al.*, 2025] delves into both explicit and implicit factors that drive emotional responses; Hi-EF [Wang *et al.*, 2026a] proposes forecasting future emotional trajectories; and MTMEUR [Hu *et al.*, 2025a] probes MLLMs' understanding of multi-turn emotional dialogues.

More recent efforts push further toward deeper emotional dynamics. OV-MERD [Lian *et al.*, 2025b] annotates samples with multi-label, open-vocabulary emotion descriptions, and MVEI [Wu *et al.*, 2026b] introduces the emotion statement judgment task, both of which incorporate the previously overlooked issue of subjectivity in evaluation. CA-MER [Han *et al.*, 2025] examines how models handle emotional conflicts across modalities, which is a practical yet underexplored problem. EmotionHallucer [Xing *et al.*, 2026] conducts an in-depth investigation of hallucinations in emotional reasoning, advancing reliability. Collectively, these datasets and benchmarks act as guiding beacons for the MER-with-LLMs paradigm, providing anchors and directions.

4 Multimodal Affective Representation

As illustrated in Fig. 5, to tackle Multimodal Affective Gap, current methods fall into two categories: one bridges intra-modality gaps by refining the encoding process, while the other handles inter-modality gaps by modeling interactions.

Perceptual Emotion Mapping. Multimodal encoders are typically pre-trained on general-purpose data, which makes them less effective when directly applied to capturing affective cues. Existing solutions are largely shared across different subtasks. An intuitive one is curating data and arranging training schedules, thereby adjusting the parameters of encoders or adapters to a more emotion-aligned direction. For instance, EmoDETective [Huang *et al.*, 2025] introduces a progressive training strategy that transitions from factual detection to emotional conclusion; ExpLLM [Lan *et al.*, 2025] constructs an Exp-CoT engine to synthesize data and improves FER through multi-turn conversations; BLSP-Emo [Wang *et al.*, 2024] first teaches models to understand emotional transcripts before incorporating audio signals; and EmoVerse [Li *et al.*, 2025a] jointly optimizes multiple tasks and ensures smooth training by controlling the task ratios.

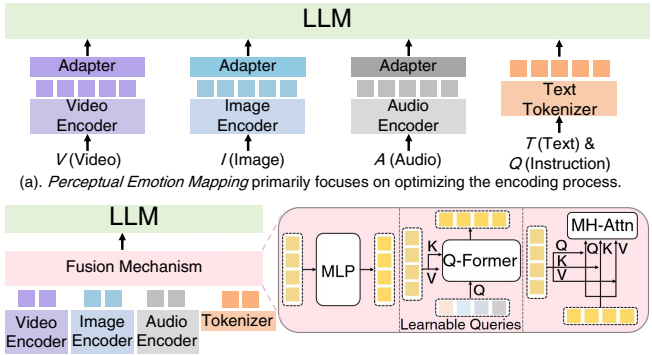

Figure 5: Differences in methodological focuses of the two types of Multimodal Affective Representation methods.

In parallel, directly modifying the encoded embeddings through sharpening or augmentation serves as another established solution. Sharpening promotes the capture of affective cues by filtering out redundant features before the LLM. For instance, SEPM [Fang *et al.*, 2025] applies coarse-to-fine inference to remove emotion-irrelevant visual tokens; Stimu-VAR [Guo *et al.*, 2025] proposes event-driven frame sampling and emotion-triggered tube selection strategies; and SE-Cap [Xu *et al.*, 2024] employs a learnable Q-former to compress audio embeddings. By contrast, augmentation complements affective cues by introducing additional features: MERMAID [Yang *et al.*, 2025c] treats MLLMs as agents and augments the original image through self-reflection; EmoLA [Li *et al.*, 2024], Emotion-LLaMA [Cheng *et al.*, 2024], FaceInsight [Li *et al.*, 2025d], and EmoChat [Xu *et al.*, 2025] implement augmentation more straightforwardly by incorporating expert encoders, with the latter two further introducing consistency constraints to facilitate feature alignment. These methods are collectively organized in Tab. 2.

Multimodal Emotion Coordination. To model the synergy of inter-modality affective cues, current works rely on modality fusion to enable thorough interactions before feeding the

	Method	Modality	Sub-task	Technique	RL Reward	Key Focus
Explanation	MulCoT-RD [Shangguan *et al.*, 2025]	I,T	VTSA	SFT	-	Computational Efficiency
	Facial-R1 [Wu *et al.*, 2026c]	I	FER	SFT+RL	Rule	Explanation Consistency
	AlignCap [Liang *et al.*, 2024]	A	SEC	SFT+RL	LLM-as-judge	Hallucination Mitigation
	EmotionReasoner [Zhao *et al.*, 2026]	V,A,T	CMER	RL	Rule	Emotional Token Weighting
	PEP-MEK [Xing *et al.*, 2026]	I,V,A,T	GVEC, CMER	Zero-shot	-	Hallucination Mitigation
	ERV [Rha *et al.*, 2026]	V,A,T	FER, CMER	SFT+RL	Model	Explanation Consistency
	FIRES [Wang *et al.*, 2025a]	V,T	MESC*	SFT+RL	Rule	Thinking Flexibility
	MultiMood [Le *et al.*, 2026]	V,A,T	CMER, MESC*	SFT+RL	LLM-as-judge	Reinforce Trustworthiness
Subject.	EmoCaliber [Wu *et al.*, 2026a]	I	GVEC	SFT+RL	Rule	Confidence Verbalization
	Agent-MER [Lai *et al.*, 2025]	V,A,T	CMER	Zero-shot	-	Open-vocabulary
	AffectGPT-R1 [Lian *et al.*, 2025c]	V,A,T	CMER	SFT+RL	Rule	Open-vocabulary

Table 4: Multimodal Affective Reasoning methods with their key attributes. MESC* stands for Multimodal Emotional Support Conversation, which takes similar inputs as CMER but focuses more on generating supportive dialogue to alleviate stress for the user.

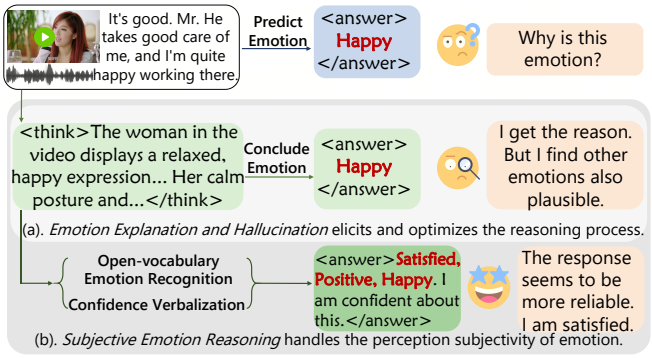

Figure 6: Evolution from sole emotion prediction, to explanation, and finally to subjectivity embracement within Multimodal Affective Reasoning methods.

representations into the LLM. We summarize these methods in Tab. 3. Overall, they primarily adopt concatenation, Q-former, and attention-based mechanisms, often accompanied by the encoder augmentation strategies discussed earlier.

Specifically, Omni-Emotion [Yang *et al.*, 2025a] finds that video-level concatenation outperforms frame-level interactions. FEALLM [Hu *et al.*, 2025b] and AffectGPT [Lian *et al.*, 2025a] employ vanilla attention mechanisms, where the former focuses on fusion across different hierarchical levels and scales, while the latter emphasizes holistic interaction. EmoVIT [Xie *et al.*, 2024] and M^{3}F [Wu *et al.*, 2025b] favor Q-former-based designs, leveraging learnable queries to extract instruction-aligned cues or retrieve memory fragments. EMO-LLaMA [Xing *et al.*, 2024] combines attention with Q-former mechanisms to inject rich facial knowledge into image embeddings. Notably, Omni-SILA [Luo *et al.*, 2025] adopts an alternative fusion paradigm based on Mixture-of-Experts for effective multi-source integration. In general, these diversified successes highlight the inherent complexity and future potential of bridging inter-modality affective gaps.

5 Multimodal Affective Reasoning

As illustrated in Fig. 6, directly predicting emotions may lead to user confusion regarding model decisions, which has motivated the emergence and development of the explainable MER task. This line of research brings diverse benefits: it not only enhances user trust but also helps uncover potential model deficiencies and optimization pathways.

Emotion Explanation and Hallucination. Methods in this branch primarily focus on the usability and reliability of explanations. To address potential inconsistencies between reasoning and predictions, Facial-R1 [Wu *et al.*, 2026c] proposes a data-centric solution, which is also supported by an RL reward that encourages factual observations. ERV [Rha *et al.*, 2026] trains a lightweight auxiliary model specifically to verify whether the answers are faithful to the reasoning. To mitigate hallucinations, PEP-MEK [Xing *et al.*, 2026] introduces a predict–explain–predict inference pipeline, inserting an additional stage to extract modality-specific and emotion-related knowledge. AlignCap [Liang *et al.*, 2024] incorporates guidance from teacher models and proprietary LLMs to enforce consistency and rationality. EmotionReasoner [Zhao *et al.*, 2026] enhances emotional interpretability by introducing an emotion-polarity token weighting mechanism that guides the reasoning process to focus on emotionally salient tokens.

Some other critical directions have also been explored. MulCoT-RD [Shangguan *et al.*, 2025] improves computational efficiency through distilling knowledge from heavy MLLMs to lightweight ones. FIRES [Wang *et al.*, 2025a] equips the model with a flexible thinking pattern that dynamically incorporates "visual scene", "emotion", "situation", and "response strategy" on demand. MultiMood [Le *et al.*, 2026] reinforces user trust in emotional support dialogue by formulating psychology-based criteria for guidance.

Subjective Emotion Reasoning. The inherent subjectivity of emotion perception allows different observers to produce different responses to the same content, posing a long-standing challenge to the MER field. Leveraging the MER-with-LLMs paradigm, several meaningful attempts have been made.

Some methods reformulate pre-defined emotion sets into open-vocabulary ones and allow multiple labels per sample, enabling models to recognize as many plausible emotions as possible. Along this direction, Agent-MER [Lai *et al.*, 2025] emulates human cognition by converting single-step emotion prediction into a structured, knowledge-driven process, achieving a balance between coverage and accuracy through self-consistent voting. AffectGPT-R1 [Lian *et al.*, 2025c] focuses on targeted reward design during RL, pro-

	Method	Taxonomy	Technique	Dataset	Metrics	Results
GVEC	SEPM [Fang *et al.*, 2025]	Perceptual Emotion Mapping	Zero-shot	EmoSet, Emotion6	Acc ↑	56.24%, 54.21%
	MERMAID [Yang *et al.*, 2025c]	Perceptual Emotion Mapping	Few-shot	EmoSet, Emotion6	Acc ↑	63.70%, 61.90%
	EmoVIT [Xie *et al.*, 2024]	Multimodal Emotion Coordination	SFT	EmoSet, Emotion6	Acc ↑	83.36%, 57.81%
VTSA	ICL [Wu *et al.*, 2025a]	Training-free Sample Configuration	Few-shot	MVSA-M	Acc ↑	69.50%
	EmoChat [Xu *et al.*, 2025]	Perceptual Emotion Mapping	SFT	MVSA-M	Acc ↑	72.70%
	MulCoT-RD [Shangguan *et al.*, 2025]	Emotion Explanation and Hallucination	SFT	MVSA-M	Acc ↑	77.20%
SEC	SECap [Xu *et al.*, 2024]	Perceptual Emotion Mapping	SFT	NNIME, EMOSEC	BLEU@4 ↑	5.8, 7.4
	AlignCap [Liang *et al.*, 2024]	Emotion Explanation and Hallucination	RL	NNIME, EMOSEC	BLEU@4 ↑	7.7, 9.8
FER	ExpLLM [Lan *et al.*, 2025]	Perceptual Emotion Mapping	SFT	RAF-DB	Acc ↑	91.03%
	EmoLA [Li *et al.*, 2024]	Perceptual Emotion Mapping	SFT	RAF-DB	Acc ↑	92.05%
	Facial-R1 [Wu *et al.*, 2026c]	Emotion Explanation and Hallucination	SFT+RL	RAF-DB	Acc ↑	92.10%
	EMO-LLaMA [Xing *et al.*, 2024]	Multimodal Emotion Coordination	SFT	DFEW	UAR ↑	60.23%
	ERV [Rha *et al.*, 2026]	Emotion Explanation and Hallucination	SFT+RL	DFEW	UAR ↑	68.88%
CMER	Emotion-LLaMA [Cheng *et al.*, 2024]	Perceptual Emotion Mapping	SFT	MER2024, MOSI, OV-MERD+	WAF ↑	73.62%, 66.13%, 52.97%
	AffectGPT [Lian *et al.*, 2025a]	Multimodal Emotion Coordination	SFT	MER2024, MOSI, OV-MERD+	WAF ↑	78.80%, 81.30%, 62.52%
	AffectGPT-R1 [Lian *et al.*, 2025c]	Subjective Emotion Reasoning	SFT+RL	MER2024, MOSI, OV-MERD+	WAF ↑	93.13%, 79.65%, 68.39%

Table 5: Quantitative performance of representative methods of MER-with-LLMs. Results are fetched from the corresponding papers. In the "Metrics" column, "Acc" denotes accuracy, "UAR" denotes unweighted average recall, and "WAF" denotes average F-score.

moting more effective optimization. Beyond reformulating the task itself, another line of work augments existing ones with a confidence dimension, allowing models to self-assess the acceptability of their predictions under potential subjective scenarios. EmoCaliber [Wu *et al.*, 2026a] progressively equips models with the ability to express and calibrate confidence, establishing a solid baseline for this direction. The methods discussed above are summarized in Tab. 4.

6 Model Performance Quantification

Following a qualitative review of methods within the MER-with-LLMs paradigm, this section presents a quantitative perspective to derive further findings and insights. As shown in Tab. 5, we observe a consistent correlation between performance and techniques across sub-tasks, revealing an evolutionary trend from freezing to fine-tuning model parameters and from SFT to RL. However, RL also introduces an observable performance fluctuation in CMER, highlighting potential room for optimization. On the other hand, methods under the Emotion Explanation and Hallucination branch tend to achieve relatively strong performance, underscoring the value of continued research into affective reasoning.

Yet these quantitative gains may obscure several underlying concerns: many existing methods still suffer from weak zero-shot generalization, limited long-tail recognition, fragile cross-cultural transfer, sentiment hallucination, and ethical risks. Although less salient than the three major challenges discussed earlier, these issues are no less important and warrant dedicated attention in future work.

7 Conclusion and Future Directions

This paper presents a comprehensive review of the emerging MER-with-LLMs paradigm. We systematically identify the key challenges and categorize existing approaches into three corresponding branches. For each branch, we discuss its motivation and representative methodologies, and compare selected works in terms of their focus, techniques, and achieved outcomes. By summarizing current progress, we aim to facilitate the future development of this paradigm. We conclude the paper by outlining the following open research directions:

Unified and Generalized MER. Existing MER research is scattered across multiple sub-tasks, all of which play indispensable roles in general emotional intelligence. Therefore, research that unifies these sub-tasks is warranted, as leveraging their complementary strengths may lead to unforeseen synergies and fuel substantial vitality.

Mechanism-level Exploration for MER. Understanding why often matters more than knowing what. The methodological progress of MER has left substantial gaps in mechanistic understanding, such as which model parameters critically influence emotion perception, and the underlying reasons why affective reasoning facilitates recognition.

Subjectivity-embraced Framework. Perception subjectivity is non-negligible in the future applications of MER systems. While greatly inspiring, current frameworks still leave substantial room for refinement. For example, it remains underexplored how to establish theoretically grounded open-vocabulary emotions, as well as how to evaluate open-ended interpretations. Alternative frameworks also merit investigation, such as determining how subjectivity should be incorporated into models, and how to design subjectivity-enriched datasets and corresponding training strategies.

Agentic Emotion Understanding. Direct interaction with the real world is an inevitable path for future MER development. To bridge perception with actionable emotional intelligence, it is crucial to establish training and evaluation frameworks that support observation, planning, reasoning, tool invocation, and real-time feedback.

Safety, Bias, and Cultural Adaptation in MER. Ensuring safety in MER requires addressing risks related to cultural variation, social bias, and individual differences. Current research largely remains at the stage of dataset construction and empirical analysis of general-purpose MLLMs, with limited exploration of concrete solutions. Advancing this area demands principled model designs and training strategies that enable culturally adaptive, bias-aware, and personalized emotion understanding for robust and responsible deployment.

Acknowledgments

This work is supported by Beijing Natural Science Foundation (No. L252009), the National Natural Science Foundation of China (Nos. 62571294, 62441614), the Open Research Fund from Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ) (No. GML-KF-26-09), and the Open Project of Xinjiang Key Laboratory of Multimodal Intelligent Computing and Large Models, Kashi University.

Contribution Statement

Hongrui Zhang and Daiqing Wu contribute equally to this work.

References

[Cai *et al.*, 2022] Qian Cai, Guo-Chong Cui, et al. Eeg-based emotion recognition using multiple kernel learning. *Machine Intelligence Research*, 2022.

[Cheng *et al.*, 2024] Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Kai Wang, Yuxiang Lin, Zheng Lian, et al. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. In *NeurIPS*, 2024.

[Chu *et al.*, 2025] Yuqi Chu, Lizi Liao, Zhiyuan Zhou, Chong-Wah Ngo, et al. Towards multimodal emotional support conversation systems. *IEEE TMM*, 2025.

[Etesam *et al.*, 2024] Yasaman Etesam, Özge Nilay Yalçın, Chuxuan Zhang, et al. Contextual emotion recognition using large vision language models. In *IROS*, 2024.

[Fang *et al.*, 2025] Yiyang Fang, Jian Liang, Wenke Huang, He Li, Kehua Su, and Mang Ye. Catch your emotion: Sharpening emotion perception in multimodal large language models. In *ICML*, 2025.

[Gao *et al.*, 2025] Lancheng Gao, Ziheng Jia, Yunhao Zeng, Wei Sun, Yiming Zhang, et al. Eemo-bench: A benchmark for multi-modal large language models on image evoked emotion assessment. In *ACM MM*, 2025.

[Guo *et al.*, 2025] Yuxiang Guo, Faizan Siddiqui, Yang Zhao, Rama Chellappa, and Shao-Yuan Lo. Stimuvar: Spatiotemporal stimuli-aware video affective reasoning with multimodal large language models. *IJCV*, 2025.

[Han *et al.*, 2025] Zhiyuan Han, Beier Zhu, Yanlong Xu, et al. Benchmarking and bridging emotion conflicts for multimodal emotion reasoning. In *ACM MM*, 2025.

[Hu *et al.*, 2025a] Jinpeng Hu, Hongchang Shi, Chongyuan Dai, Zhuo Li, Peipei Song, and Meng Wang. Beyond emotion recognition: A multi-turn multimodal emotion understanding and reasoning benchmark. In *ACM MM*, 2025.

[Hu *et al.*, 2025b] Zhuozhao Hu, Kaishen Yuan, Xin Liu, Zitong Yu, Yuan Zong, et al. Feallm: Advancing facial emotion analysis in multimodal large language models with emotional synergy and reasoning. In *ACM MM*, 2025.

[Huang *et al.*, 2025] Xuandong Huang, Yuzhe Zhou, Jiashu Li, et al. Emodetective: Detecting, exploring, and thinking emotional cause in videos. In *ACM MM*, 2025.

[Lai *et al.*, 2025] Zhengqin Lai, Zhilin Zhu, Xiaopeng Hong, and Yaowei Wang. Agent-mer: A cognitive agent with hierarchical deliberation for open-vocabulary multimodal emotion recognition. In *ACM MM*, 2025.

[Lan *et al.*, 2025] Xing Lan, Jian Xue, Ji Qi, Dongmei Jiang, Ke Lu, et al. Expllm: Towards chain of thought for facial expression recognition. *IEEE TMM*, 2025.

[Le *et al.*, 2026] Huy M. Le, Dat Tien Nguyen, Ngan T. T. Vo, Tuan D. Q. Nguyen, Nguyen Binh Le, Duy Minh Ho Nguyen, et al. Reinforcing trustworthiness in multimodal emotional support systems. In *AAAI*, 2026.

[Li *et al.*, 2024] Yifan Li, Anh Dao, Wentao Bao, Zhen Tan, Tianlong Chen, Huan Liu, and Yu Kong. Facial affective behavior analysis with instruction tuning. In *ECCV*, 2024.

[Li *et al.*, 2025a] Ao Li, Longwei Xu, Chen Ling, Jinghui Zhang, and Pengwei Wang. Emoverse: Enhancing multimodal large language models for affective computing via multitask learning. *Neurocomputing*, 2025.

[Li *et al.*, 2025b] Deng Li, Bohao Xing, Xin Liu, Baiqiang Xia, Bihan Wen, et al. Deemo: De-identity multimodal emotion recognition and reasoning. In *ACM MM*, 2025.

[Li *et al.*, 2025c] Jie Li, Shifei Ding, et al. Multi-modal anchor gated transformer with knowledge distillation for emotion recognition in conversation. In *IJCAI*, 2025.

[Li *et al.*, 2025d] Jingzhi Li, Changjiang Luo, Ruoyu Chen, Hua Zhang, et al. Faceinsight: A multimodal large language model for face perception. In *ACM MM*, 2025.

[Lian *et al.*, 2024a] Zheng Lian, Haiyang Sun, Licai Sun, Hao Gu, et al. Explainable multimodal emotion recognition. arXiv preprint arXiv:2306.15401, 2024.

[Lian *et al.*, 2024b] Zheng Lian, Licai Sun, Haiyang Sun, Kang Chen, Zhuofan Wen, Hao Gu, Bin Liu, and Jianhua Tao. Gpt-4v with emotion: A zero-shot benchmark for generalized emotion recognition. *INFFUS*, 2024.

[Lian *et al.*, 2025a] Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, et al. AffectGPT: A new dataset, model, and benchmark for emotion understanding with multimodal large language models. In *ICML*, 2025.

[Lian *et al.*, 2025b] Zheng Lian, Haiyang Sun, Licai Sun, Haoyu Chen, et al. OV-MER: Towards open-vocabulary multimodal emotion recognition. In *ICML*, 2025.

[Lian *et al.*, 2025c] Zheng Lian, Fan Zhang, Yazhou Zhang, Jianhua Tao, et al. Affectgpt-r1: Leveraging reinforcement learning for open-vocabulary multimodal emotion recognition. arXiv preprint arXiv:2508.01318, 2025.

[Liang *et al.*, 2024] Ziqi Liang, Haoxiang Shi, and Hanhui Chen. AlignCap: Aligning speech emotion captioning to human preferences. In *EMNLP*, 2024.

[Lin *et al.*, 2025] Yuxiang Lin, Jingdong Sun, Zhi-Qi Cheng, Jue Wang, Haomin Liang, et al. Why we feel: Breaking boundaries in emotional reasoning with multimodal large language models. In *CVPR Workshops*, 2025.

[Luo *et al.*, 2025] Jiamin Luo, Jingjing Wang, Junxiao Ma, Yujie Jin, Shoushan Li, and Guodong Zhou. Omni-sila:

Towards omni-scene driven visual sentiment identifying, locating and attributing in videos. In *WWW*, 2025.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[Rha *et al.*, 2026] Hyeongseop Rha, Jeong Hun Yeo, Yeonju Kim, et al. Emotion-coherent reasoning for multimodal llms via emotional rationale verifier. In *AAAI*, 2026.

[Shangguan *et al.*, 2025] Haonan Shangguan, Xiaocui Yang, Shi Feng, Daling Wang, Yifei Zhang, and Ge Yu. Resource-limited joint multimodal sentiment reasoning and classification via chain-of-thought enhancement and distillation. arXiv preprint arXiv:2508.05234, 2025.

[Shou *et al.*, 2025] Yuntao Shou, Tao Meng, Wei Ai, et al. Multimodal large language models meet multimodal emotion recognition and reasoning: A survey. arXiv preprint arXiv:2509.24322, 2025.

[Wang *et al.*, 2024] Chen Wang, Minpeng Liao, Zhongqiang Huang, Junhong Wu, et al. BLSP-emo: Towards empathetic large speech-language models. In *EMNLP*, 2024.

[Wang *et al.*, 2025a] Fanfan Wang, Xiangqing Shen, Jianfei Yu, and Rui Xia. Flexible thinking for multimodal emotional support conversation via reinforcement learning. In *Findings of EMNLP*, 2025.

[Wang *et al.*, 2025b] Xiaoqin Wang, Xusen Ma, Xianxu Hou, Meidan Ding, Yudong Li, et al. Facebench: A multi-view multi-level facial attribute vqa dataset for benchmarking face perception mllms. In *CVPR*, 2025.

[Wang *et al.*, 2025c] Zhifeng Wang, Qixuan Zhang, Peter Zhang, et al. Visual and textual prompts in vllms for enhancing emotion recognition. *IEEE TCSVT*, 2025.

[Wang *et al.*, 2026a] Haoran Wang, Xinji Mai, Zeng Tao, Junxiong Lin, Xuan Tong, et al. Hi-ef: Benchmarking emotion forecasting in human-interaction. In *AAAI*, 2026.

[Wang *et al.*, 2026b] Taorui Wang, Xun Lin, Yong Xu, et al. Micro-gesture recognition: A comprehensive survey of datasets, methods, and challenges. *Machine Intelligence Research*, 2026.

[Wu *et al.*, 2025a] Daiqing Wu, Dongbao Yang, Sicheng Zhao, et al. An empirical study on configuring in-context learning demonstrations for unleashing MLLMs' sentimental perception capability. In *ICML*, 2025.

[Wu *et al.*, 2025b] Dan Wu, Xincheng Ju, Dong Zhang, Shoushan Li, et al. Emotion across modalities and cultures: Multilingual multimodal emotion-cause analysis with memory-inspired framework. In *ACM MM*, 2025.

[Wu *et al.*, 2026a] Daiqing Wu, Dongbao Yang, Can Ma, and Yu Zhou. Emocaliber: Advancing reliable visual emotion comprehension via confidence verbalization and calibration. *PR*, 2026.

[Wu *et al.*, 2026b] Daiqing Wu, Dongbao Yang, Sicheng Zhao, et al. Customizing visual emotion evaluation for mllms: An open-vocabulary, multifaceted, and scalable approach. In *ICLR*, 2026.

[Wu *et al.*, 2026c] Jiulong Wu, Yucheng Shen, Lingyong Yan, Haixin Sun, et al. Facial-r1: Aligning reasoning and recognition for facial emotion analysis. In *AAAI*, 2026.

[Xie *et al.*, 2024] Hongxia Xie, Chu-Jun Peng, Yu-Wen Tseng, et al. Emovit: Revolutionizing emotion insights with visual instruction tuning. In *CVPR*, 2024.

[Xing *et al.*, 2024] Bohao Xing, Zitong Yu, Xin Liu, Kaishen Yuan, Qilang Ye, et al. Emo-llama: Enhancing facial emotion understanding with instruction tuning. arXiv preprint arXiv:2408.11424, 2024.

[Xing *et al.*, 2026] Bohao Xing, Xin Liu, Guoying Zhao, et al. Emotionhallucer: Evaluating emotion hallucinations in multimodal large language models. In *ICLR*, 2026.

[Xu *et al.*, 2024] Yaoxun Xu, Hangting Chen, Jianwei Yu, Qiaochu Huang, Zhiyong Wu, et al. Secap: Speech emotion captioning with large language model. In *AAAI*, 2024.

[Xu *et al.*, 2025] Qinfu Xu, Liyuan Pan, Shaozu Yuan, Yiwei Wei, and Chunlei Wu. From subtle hints to grand expressions - mastering fine-grained emotions with dynamic multimodal analysis. In *ACM MM*, 2025.

[Yang *et al.*, 2023] Xiaocui Yang, Wenfang Wu, Shi Feng, Ming Wang, Daling Wang, et al. Mm-bigbench: Evaluating multimodal models on multimodal content comprehension tasks. arXiv preprint arXiv:2310.09036, 2023.

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

[Yang *et al.*, 2025b] Xudong Yang, Yizhang Zhu, Hanfeng Liu, Zeyi Wen, et al. Ramer: Reconstruction-based adversarial model for multi-party multi-modal multi-label emotion recognition. In *IJCAI*, 2025.

[Yang *et al.*, 2025c] Zhongyu Yang, Junhao Song, et al. MERMAID: Multi-perspective self-reflective agents with generative augmentation for emotion recognition. In *EMNLP*, 2025.

[Zhang *et al.*, 2024] Qixuan Zhang, Zhifeng Wang, Dylan Zhang, Wenjia Niu, et al. Visual prompting in LLMs for enhancing emotion recognition. In *EMNLP*, 2024.

[Zhao *et al.*, 2021a] Sicheng Zhao, Guoli Jia, Jufeng Yang, et al. Emotion recognition from multiple modalities: Fundamentals and methodologies. *IEEE SPM*, 2021.

[Zhao *et al.*, 2021b] Sicheng Zhao, Xingxu Yao, Jufeng Yang, et al. Affective image content analysis: Two decades review and new perspectives. *IEEE TPAMI*, 2021.

[Zhao *et al.*, 2023] Sicheng Zhao, Xiaopeng Hong, Jufeng Yang, Yanyan Zhao, and Guiguang Ding. Toward label-efficient emotion and sentiment analysis. *PIEEE*, 2023.

[Zhao *et al.*, 2026] Sicheng Zhao, Hongrui Zhang, Daiqing Wu, et al. Emotionreasoner: Emotion-explanation-oriented reinforcement learning for explainable multimodal emotion recognition. *IEEE TAFFC*, 2026.