Listen, Watch, and Learn to Feel: Retrieval-Augmented Emotion Reasoning for Compound Emotion Generation

Zhuofan Wen[1,2] **Zheng Lian**[2] **Shun Chen**[1,2] **Hailiang Yao**[1,2] **Longjiang Yang**[1,2]
Bin Liu[1,2] * **Jianhua Tao**[3,4]

[1]School of Artificial Intelligence, University of Chinese Academy of Sciences,
[2]Institute of Automation, Chinese Academy of Sciences,
[3]The Department of Automation, Tsinghua University
[4]BNRIST, Tsinghua University

wenzhuofan2023@ia.ac.cn liubin@nlpr.ia.ac.cn

Abstract

The ability to comprehend human emotion using multimodal large language models (MLLMs) is essential for advancing human-AI interaction and multimodal sentiment analysis. While psychology theory-based human annotations have contributed to multimodal emotion tasks, the subjective nature of emotional perception often leads to inconsistent annotations, limiting the robustness of current models. Addressing these challenges requires more fine-grained methods and evaluation frameworks. In this paper, we propose the Retrieval-Augmented Emotion Reasoning (RAER) framework, a plug-and-play module that enhances MLLMs' ability to tackle compound and context-rich emotion tasks. To systematically evaluate model performance, we introduce the Stimulus-Armed Bandit (SAB) framework, designed to benchmark emotional reasoning capabilities. Additionally, we construct the Compound Emotion QA dataset, an AI-generated multimodal dataset aimed at strengthening emotion understanding in MLLMs. Experimental results demonstrate the effectiveness of RAER across both traditional benchmarks and SAB evaluations, highlighting its potential to enhance emotional intelligence in multimodal AI systems.

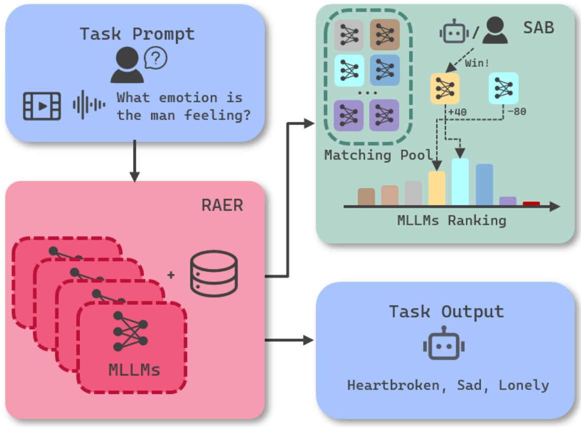


Figure 1: Overall of our proposed method and evaluation framework, where the Retrieval-Augmented Emotion Reasoning (RAER) method is introduced as a plug-and-play module to enhance the capability of MLLMs in handling compound and ambiguous emotions. The Stimulus-Armed Bandit (SAB) evaluation framework is used to assess the model's emotional capabilities, especially for tasks that are difficult to quantify.

1 Introduction

Emotion is a multifaceted phenomenon encompassing subjective experiences, physiological responses, and context-dependent behaviors, shaped by both internal states and external stimuli. Emotions play a critical role in human cognition and interaction, influencing decision-making, directing attention, and shaping social relationships. Their complexity and impact underscore the significance of emotions as a core component of human experience and behavior.

Recent advancements in neural network methods have highlighted the effectiveness of specialized models for emotion tasks. These models, which predict labels within a constrained range, have achieved impressive performance, particularly in tasks such as Dynamic Facial Emotion Recognition (DFER) (Tran et al., 2015; Wang et al., 2023; Ghaleb et al., 2019) and Multimodal Emotion Recognition (MER) (Tsai et al., 2019; Hazarika et al., 2020; Zadeh et al., 2018).

However, widely adopted annotation standards within a constrained range—such as the "Big Six" discrete label system (Ekman, 1992) and the VAD (Valence-Arousal-Dominance) dimensional label system (Russell and Mehrabian, 1977)—have proven effective in capturing emotional expression, they may not fully align with the more nuanced emotional interactions required for AI systems, particularly in the era of large models that demand more human-like interaction for emotion-related tasks. To address these limitations, ap-

* Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2025, pages 11313–11327
July 27 - August 1, 2025

proaches based on multimodal large language models (MLLMs) have emerged (Lian et al., 2024c; Cheng et al., 2024a). Tasks such as Multimodal Empathetic Response Generation (MERG) (Zhang et al., 2024a) and other emotion-related tasks (Zheng et al., 2024; Sabour et al., 2024; Plaza Del Arco et al., 2024) have shown strong performance on MLLMs, with excellent generalization capabilities. Despite these advancements, significant challenges remain in handling compound and ambiguous emotions, especially in tasks involving compound and context-rich emotional scenarios.

In this paper, as shown in Figure 2, we draw inspiration from recent advances in preference-based learning methods—such as Reinforcement Learning with Human Feedback (RLHF) (Stiennon et al., 2022), Reinforcement Learning from AI Feedback (RLAIF) (Bai et al., 2022), Direct Preference Optimization (DPO) (Rafailov et al., 2024), and Inverse Preference Optimization (IPO) (Huang et al., 2024b)—to explore a more fluid and subjective approach to evaluating emotion tasks. However, a major challenge in constructing emotion preference datasets lies in the labor-intensive process of manually drafting labels, which is not only time-consuming but also prone to inconsistencies. These inconsistencies arise from both variations in linguistic descriptions and differences in human preferences, making it difficult to disentangle the specific preference signals required for AI model training (Lian et al., 2024a; Cheng et al., 2024a).

As shown in Figure.1. Instead of relying on manually curated label drafts and preference annotations, we propose a Retrieval-Augmented Emotion Reasoning (RAER) Framework, a RAG-based module that integrates chain-of-thought (CoT) reasoning. RAER can be easily applied to MLLMs, enhancing their emotional reasoning and generalization capabilities to tackle compound emotional scenarios. To evaluate MLLMs' emotion task capabilities and gather human preferences, we introduce the Stimulus-Armed Bandit (SAB) Evaluation Framework, which uses AI-generated stimuli to test a broad range of emotion tasks. This approach not only collects human preferences but also benchmarks model performance in dynamic and compound emotional contexts. By combining RAER-generated responses with SAB-collected human preferences, we construct the Compound Emotion QA Dataset, a multimodal dataset that captures nuanced emotional reasoning aligned with human preferences. This methodology bridges the gap between traditional emotion recognition and more compound emotional reasoning, offering a scalable and preference-aligned solution to advance emotional intelligence in MLLMs.

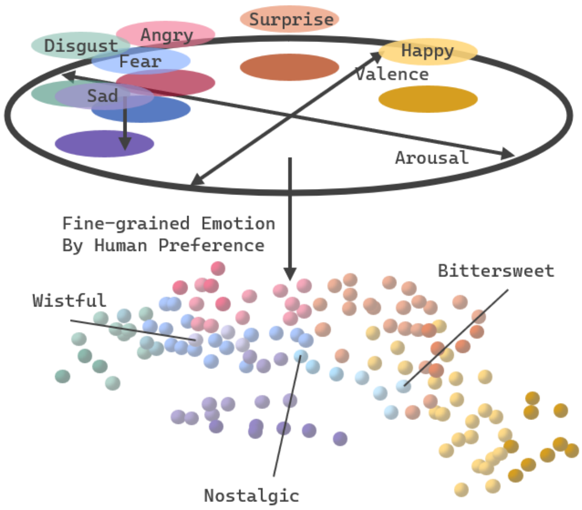


Figure 2: In linguistic contexts, the expression of human emotions is inherently open-ended, suggesting that label systems with predefined boundaries are limited in the context of large models. To enhance the emotional capabilities of these models, it is essential to adopt more human-like, nuanced approaches that allow for a broader range of emotional expression.

Main Contributions. The major contributions of this work are summarized as follows:

- Emotion Reasoning RAG: We propose a retrieval-augmented framework, Retrieval-Augmented Emotion Reasoning (RAER), which incorporates a chain-of-emotion reasoning approach to enhance MLLMs' capability in addressing compound emotion tasks.
- Stimulus-Armed Bandit (SAB) Evaluation Framework: We introduce the SAB framework to systematically evaluate MLLMs' performance in compound emotional scenarios.
- Compound Emotion QA: We construct a multimodal QA dataset that includes compound emotion tasks, designed to enhance the compound emotional capabilities of MLLMs.

2 Related Work

2.1 Multimodal Emotion Recognition

Multimodal Emotion Recognition (MER) aims to improve emotion detection by integrating multiple modalities. With the rise of neural network-based methods, advanced modality fusion networks have been proposed (Tsai et al., 2019; Hazarika et al., 2020; Zadeh et al., 2018), leading to significant improvements in MER. However, challenges such as high dataset labeling costs and task-specific network architectures still hinder the models' generalization in real-world scenarios.

The development of MLLMs has shown promising improvements in MER. These models, leveraging large-scale pretraining on diverse multimodal data, demonstrate enhanced generalization capabilities in emotion tasks (Lian et al., 2024c; Cheng et al., 2024a), surpassing traditional approaches in terms of flexibility and adaptability.

Moreover, the ongoing expansion of multimodal emotion recognition benchmarks and datasets has contributed significantly to this progress (Sabour et al., 2024; Lian et al., 2024c). These resources allow for more comprehensive evaluations of MLM-based models, supporting more accurate emotion detection and better alignment with real-world applications .

2.2 Multimodal Empathetic Response Generation

Multimodal empathetic response generation aims to enable machines to not only understand human emotions but also respond empathetically across various modalities. While Large Language Models (LLMs) (Zhang et al., 2024a; Yang et al., 2024), have shown potential in generating empathetic responses from textual inputs, incorporating additional modalities such as voice tone, facial expressions, and body language remains an ongoing challenge. Furthermore, the subjective nature of empathy complicates the evaluation process, making it difficult to define consistent and reliable metrics for assessing the quality of empathetic responses (Wu et al., 2024). These challenges emphasize the need for further research into better integration of multimodal data to enhance the emotional depth and reliability of empathetic response generation systems.

2.3 Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) has proven effective in enhancing generative models' ability to generalize across tasks by incorporating external knowledge, such as in Text-to-3D (Seo et al., 2024) and Protein Molecule Generation (Huang et al., 2024d). In emotion-related tasks, RAG has been used to improve response generation by dynamically retrieving emotionally relevant data to refine models' outputs (Huang et al., 2024a; Liu et al., 2024). These approaches have been applied in areas like emotional agent (Huang et al., 2024a) and Empathetic response generation (ERC) (Huang et al., 2024c), where diverse emotional cues enhance performance.

Building on these methods, our work extends RAG to compound emotion tasks by incorporating contextual emotional knowledge from multimodal sources. This approach improves emotion recognition and generation by using dynamic, context-driven retrieval, enabling more flexible and empathetic models that can handle a wider range of emotional scenarios.

3 Retrieval-Augmented Emotion Reasoning

Emotional reasoning refers to the process of deriving conclusions based on emotional responses, even when empirical evidence may suggest otherwise. This concept has proven effective in large models for addressing compound emotion-related tasks (Lian et al., 2023), particularly those involving ambiguous or context-dependent emotional content. Building on this foundation, as shown in Figure 3, we propose Retrieval-Augmented Emotion Reasoning (RAER), a framework designed to enhance multimodal large language models (MLLMs) by integrating emotional reasoning into a structured chain-of-thought (CoT) process (Wei et al., 2023).

3.1 Building the Emotional Knowledge Base

A cornerstone of RAER is the emotional knowledge base, which serves as the foundation for retrieval during the reasoning process. Initially, the knowledge base is constructed from multimodal emotion datasets, encoding diverse inputs such as facial expression animations, emotional audio clips, and human/AI-generated emotional descriptions. Each sample is transformed into a high-dimensional vector embedding, enriched with detailed emotional annotations, enabling efficient similarity-based retrieval to support the reasoning process. As RAER engages in iterative reasoning tasks, the knowledge base evolves through the addition of high-confidence samples generated during the reasoning process. This dynamic updating mechanism not only enhances the diversity of the knowledge base but also improves its ability

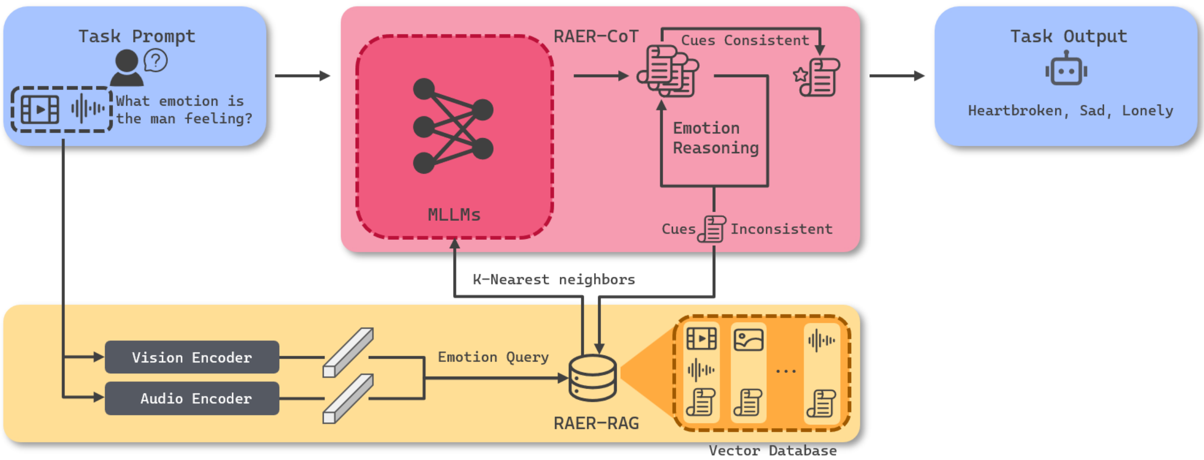


Figure 3: Overall of the RAER architecture, where a multimodal prompt is input into the model. RAER searches for the most similar content in the vector database, incorporating the Emotion Reasoning CoT process to validate the emotional consistency of the model's response. After ensuring emotional coherence, the model generates the final output.

to provide contextually relevant emotional references. By combining curated data from traditional datasets with newly derived samples, the emotional knowledge base becomes a continuously expanding resource, progressively strengthening RAER's capacity for compound and nuanced emotional reasoning.

3.2 Guiding Emotion Reasoning with Chain-of-Thought

The RAER framework leverages a CoT reasoning mechanism to guide MLLMs through emotional reasoning tasks. This structured approach allows models to process compound emotional inputs step by step, identifying uncertainties or ambiguities within generated captions or descriptions. These challenges often arise in scenarios involving overlapping or conflicting emotional cues. To address them, RAER incorporates retrieval-based augmentation, enabling models to draw upon contextually relevant emotional information retrieved from an external knowledge base.

3.3 Enhancing Emotional Reasoning with Retrieval

Incorporating retrieval into the reasoning process allows RAG to refine the model's understanding of compound emotions. For example, when a caption reflects emotional ambiguity, the framework retrieves similar examples from the emotional knowledge base, along with their associated emotional descriptions. By grounding its reasoning in these examples, RAER enables the model to refine its understanding, disambiguate emotional cues, and generate more accurate and contextually appropriate inferences.

Algorithm 1 Retrieval-Augmented Emotion Reasoning (RAER)

Input: Task prompt I, multimodal input X (e.g., video, audio, text), knowledge base $\mathcal{K}$, MLLM f_θ
Output: Refined reasoning outputs $\{y_t\}_{t=1}^T$

Generate initial response: $Y = f_\theta(I, X)$
Segment reasoning steps: $\{y_i\}_{i=1}^T = \text{Segment}(f_{\text{Analyze}}(Y))$
for $t = 1$ to T **do**
    Retrieve context: $R(y_{t-1}) = \text{Retrieve}(\mathcal{K}, \text{Sim}(y_{t-1}, \mathcal{K}))$
    if Detect$(y_t, R(y_{t-1}))$ is ambiguous **then**
        Generate reasoning step: $y_t = f_{\text{CoT}}(y_t, R(y_t))$
    end if
end for
if Uncertainty$(y_t) \leq \epsilon$ **then**
    Update knowledge base: $\mathcal{K} \leftarrow \mathcal{K} \cup \{(X, \{y_t\}_{t=1}^T)\}$
end if
Return: Refined reasoning steps $\{y_t\}_{t=1}^T$

Through RAER, we enhance MLLMs' zero-shot capabilities for handling emotion tasks. However, to further improve model performance with human feedback, we need precise human preference signals and a method to evaluate generative models' performance on emotion tasks in open-ended language contexts. To address this, we designed the

Stimulus-Armed Bandit(SAB) framework.

4 Stimulus-Armed Bandit

The Stimulus-Armed Bandit (SAB) framework is a novel evaluation method designed to assess the compound emotion capabilities of multimodal large language models (MLLMs). The name is inspired by the classic multi-armed bandit optimization problem, as emotion tasks similarly involve balancing between different emotional responses. SAB introduces a preference-based ranking system that combines multimodal stimuli with emotion tasks. Through pairwise comparisons, SAB dynamically ranks models based on their emotional reasoning, while also collecting human preferences and corresponding task labels.

The SAB framework integrates three key components to enable comprehensive and scalable evaluation:

(1) Stimulus Generation, which produces multimodal triggers to elicit emotional responses;

(2) Task Formulation, which combines stimuli with emotion-related downstream tasks to create diverse evaluation challenges; and

(3) Ranking Mechanism, which utilizes an elo-based scoring system to dynamically adjust model rankings based on their comparative performance.

4.1 Stimulus Generation and Task Formulation

Stimulus Generation. Stimuli serve as the core of the SAB framework, functioning as controlled triggers designed to evoke specific emotional and cognitive responses. To achieve this, single or multiple emotion-neutral keywords are randomly sampled, and LLMs are prompted to improvisationally generate human-centered scenario prompts. Generative models then create corresponding content based on these prompts. This approach ensures that the generated content aligns with real-world emotional scenarios rather than pre-defined emotional contexts. Leveraging AI-generated content, stimuli are dynamically delivered with diverse and non-repetitive samples across multiple modalities, including text, audio, images, and video, providing a broad spectrum of emotional triggers for evaluation.

Task Formulation. Each stimulus is paired with a corresponding downstream task commonly seen in MER or MERG, These tasks are randomly assigned to MLLMs to ensure diverse and unbiased evaluations. The randomized pairing of stimuli and tasks ensures that MLLMs are exposed to a wide range of scenarios, testing their adaptability and robustness in compound emotional contexts.

4.2 Ranking Mechanism

Initialization. All MLLMs start with identical ranking scores, ensuring a fair initial condition. The starting score is set to a predefined baseline, S_0, which is the same for all participating models.

Pairwise Matching and Preference Judgments. During each evaluation round, MLLMs are first paired based on similar ranking scores to ensure competitive and balanced matches. Let the models be denoted as $M_1, M_2, \dots, M_N$, where each M_i has a score S_i. The models are paired such that $|S_i - S_j|$ is minimized for each match.Next, each pair of models is assigned identical stimuli and tasks drawn randomly from the task pool. For instance, if a pair M_i and M_j is matched, they both receive the same stimulus X_{stimulus} and task T_{task}. Once the stimuli and tasks are assigned, each model generates a response to the task, denoted as R_i for model M_i and R_j for model M_j. The responses R_i and R_j are then evaluated by human or AI evaluators, who compare them based on criteria such as emotional relevance, coherence, and depth of reasoning. Finally, the evaluators select the preferred response, which could be denoted as $R_{\text{preferred}}$, where:

$$R_{\text{preferred}} = \begin{cases} R_i, & \text{if model } M_i \text{ is preferred,} \\ R_j, & \text{if model } M_j \text{ is preferred.} \end{cases}$$

This process ensures a robust and contextually relevant evaluation based on human-like judgments or AI preferences.

Score Adjustment. The ranking scores are updated using an elo-based mechanism, as follows:

$$S_i^{\text{new}} = S_i^{\text{old}} + K \cdot (R - E), \quad (1)$$

where: S_i^{new} denotes the updated ranking score of model i, S_i^{old} represents the current ranking score of model i, K is the scaling factor that determines the sensitivity of score adjustments, M is a tunable parameter that controls the ranking sensitivity, R indicates the actual match outcome, where $R = 1$ if the model wins, $R = 0.5$ for a draw, and $R = 0$ if the model loses, E represents the expected match outcome, calculated as:

$$E = \frac{1}{1 + 10^{(S_j - S_i)/M}}, \quad (2)$$

where S_j is the opponent's ranking score.

Iterative Refinement. Over multiple rounds, models compete against opponents with similar scores, gradually revealing their relative strengths. This mechanism ensures fair and adaptive ranking, reflecting each model's ability to handle compound emotion tasks.

5 Experiments

In this section, we present a comprehensive evaluation of the Retrieval-Augmented Emotion Reasoning (RAER) framework. The experiments are conducted across two main categories: Traditional datasets and metrics, encompassing various tasks in Multimodal Emotion Recognition (MER) and Multimodal Empathetic Response Generation (MERG), and the novel Stimulus-Armed Bandit (SAB) evaluation framework, which systematically evaluates both visual-language and audio-language generated samples.

5.1 Implementation Details

MLLMs Implementation. In our experiments, the models under consideration include general-purpose MLLMs—VideoLLaMA2 (Cheng et al., 2024b), LlavaNextVideo (Zhang et al., 2024b), Qwen2.5-VL (Team, 2025), SALMONN (Sun et al., 2024), and Qwen2-Audio (Chu et al., 2024)—as well as emotion-task-specific fine-tuned MLLMs: AffectGPT (Lian et al., 2024c) and EmotionLLaMA (Cheng et al., 2024a). The limitations of supervised fine-tuning (SFT) for current emotion tasks are discussed in detail in Appendix D . With the exception of AffectGPT and EmotionLlama, none of the models were fine-tuned on emotion tasks. During the course of the experiments, we find that AffectGPT and EmotionLlama support a maximum context size of 2048 tokens, which lead to failures in RAER's CoT process due to the limited context length of the models. As a result, we decide not to incorporate the RAER module into these models. All models in our experiments used a 7B parameter size. WAR (Weighted Average Recall) is employed as the evaluation metric across all datasets in MER Task Evalution, and ablation studies are conducted to examine the integration of RAER.

RAER Implementation. We use Faiss (Douze et al., 2024) as the vector dataset for RAER, employing SBERT (Reimers and Gurevych, 2019) for encoding the textual content and CLIP (Radford et al., 2021) for encoding the visual content. The features from multiple frames are concatenated to form the query vector. For the audio content, we use AST (Gong et al., 2021) for encoding. The initial samples are derived from the 332 labeled entries in the EMER-finev2 (Lian et al., 2024a) dataset. As inference progresses on each dataset and RAER continues to augment its vectorbase, we observe that the model benefits from increasingly diverse and representative retrieval samples. As shown in Figure 4, we evaluated this effect on MER2024, DFEW, and IEMOCAP. At each RAER epoch, the model performs inference over the entire dataset, and the retrieval database is dynamically expanded by incorporating newly inferred samples with high-confidence and accurate predictions. In the subsequent Traditional Benchmarks Evaluation, we utilize the RAER vector database obtained after the sixth epoch of retrieval-augmented inference on each dataset. Table 1 summarizes the number of high-confidence samples added by RAER after six inference epochs on each dataset. For the following SAB Evaluation, we construct the RAER retrieval vector database by merging the newly added samples from all three benchmark datasets with the initial 332 samples from EMER-finev2, thereby forming a more comprehensive and diverse retrieval foundation.

Dataset	Total Samples	RAER-Added
MER2024	5,030	279
DFEW	11,697	958
IEMOCAP	10,039	351

Table 1: Number of high-confidence samples added by RAER after six inference epochs.

SAB Evaluation Setup. For the SAB evaluation, we use GPT-4 to generate task prompts, OpenAI's Sora for visual-language evaluation, and Meta's AudioGen (Kreuk et al., 2022) for audio-language evaluation. As no existing generative model is capable of simultaneously producing visual, audio, and textual modalities, we are unable to perform SAB evaluation on VAT(visual-audio-text) samples.

The experiments are conducted using four NVIDIA H800 80GB GPUs and The hyperparameters used for RAER and SAB are detailed in Appendix A..

5.2 Traditional Benchmarks Evaluation

MER Tasks Evaluation. As shown in Table 9. We conduct a comprehensive evaluation of RAER's impact on the multimodal emotion recognition capa-

MLLMs	V	A	T	MER2024 w/o RAER	MER2024 RAER	DFEW w/o RAER	DFEW RAER	IEMOCAP w/o RAER	IEMOCAP RAER
AffectGPT	✓	✓	✓	**0.64**	-	0.52	-	**0.71**	-
EmotionLlama	✓	✓	✓	0.23	-	**0.59**	-	0.32	-
VideoLlama2	✓	✓	✓	0.35	0.62	0.32	0.66	0.41	0.71
LlavaNextVideo	✓	✗	✓	0.23	0.37	0.22	0.32	0.28	0.40
Qwen2.5VL	✓	✗	✓	0.45	**0.73**	0.43	**0.69**	0.53	**0.78**
SALMONN	✓	✓	✗	0.31	0.42	0.24	0.37	0.34	0.39
Qwen2Audio	✓	✓	✗	0.27	0.35	0.22	0.28	0.31	0.37

Table 2: The results of MER tasks are presented, where WAR is used as the evaluation metric. All results are obtained using zero-shot inference.

MLLMs	BLEU w/o RAER	BLEU RAER	Dist-1/2 w/o RAER	Dist-1/2 RAER	ROU_L. w/o RAER	ROU_L. RAER	MET. w/o RAER	MET. RAER	BERTS. w/o RAER	BERTS. RAER
AffectGPT	0.13	-	0.56/0.82	-	0.17	-	0.31	-	**0.87**	-
EmotionLlama	0.10	-	0.54/0.79	-	0.15	-	0.28	-	0.85	-
VideoLlama2	**0.22**	0.24	0.71/0.92	0.69/0.92	**0.23**	0.24	0.31	**0.40**	0.83	0.87
LlavaNextVideo	0.17	0.22	0.69/0.87	0.71/0.86	0.19	0.2	0.28	0.25	0.79	0.83
Qwen2.5VL	0.21	**0.25**	**0.74/0.95**	**0.77/0.95**	**0.23**	**0.25**	**0.33**	0.36	0.86	**0.91**
SALMONN	0.18	0.17	0.72/0.88	0.72/0.89	0.17	0.21	0.23	0.25	0.81	0.85
Qwen2Audio	0.15	0.22	0.68/0.85	0.69/0.87	0.18	0.22	0.25	0.27	0.83	0.86

Table 3: The results of the automatic evaluation of MERG tasks are presented, with all outcomes obtained through zero-shot inference.

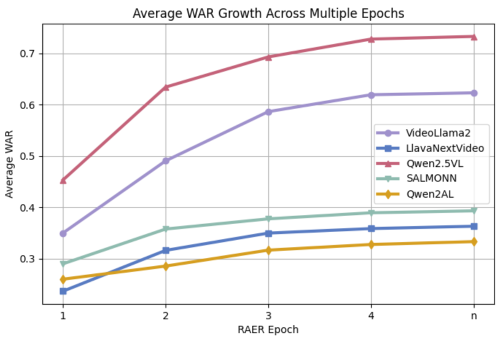
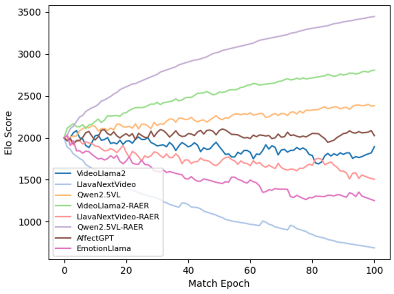


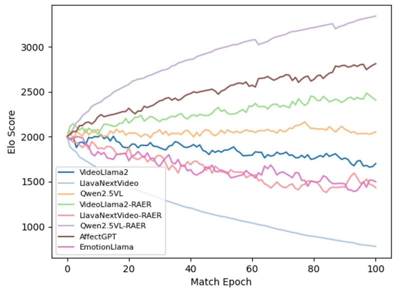
Figure 4: In the RAER framework, correctly predicted samples from the dataset are added to the vector database and referenced in subsequent predictions. After multiple rounds of this process, the model demonstrates significant performance improvements, as shown by the average WAR growth across epochs.

bilities of various multimodal large language models (MLLMs) through ablation studies. Detailed results for specific metrics, including per-emotion classification accuracy and the unweighted average recall (UAR) for each dataset, are provided in Appendix B. Models are evaluated on three widely recognized multimodal emotion recognition (MER) datasets: MER2024 (Lian et al., 2024b), DFEW (Jiang et al., 2020), and IEMOCAP (Busso et al., 2008). In our experiments, Qwen2.5VL demonstrate remarkable performance. Despite the absence of audio modality information, Qwen2.5VL-RAER still achieve a WAR score of over 0.7 across three datasets, outperforming both AffectGPT and EmotionLlama, which had been fine-tuned on the MER task in a zero-shot setting. Additionally, we observed that visual information contributed more significantly to performance improvements compared to audio information.The experimental results consistently demonstrate that REAR significantly enhances the emotion recognition performance of these MLLMs, highlighting its robust ability to improve emotional comprehension across diverse multimodal contexts.

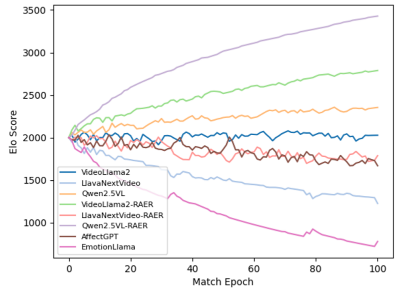
MERG Tasks Evaluation. As shown in Table 3, We first conducted automated evaluations using several standard metrics, including BLEU (Papineni et al., 2002), Dist-1/2 (Li et al., 2016), ROUGE-L (Lin, 2004), METEOR (Banerjee and Lavie, 2005), and BERTScore (Zhang et al., 2020). Following the precious approaches (Wu et al., 2025; Zhang et al., 2024a; Fei et al., 2024), we also designed human evaluation protocols to assess various aspects such as response empathy, linguistic fluency, and consistency in Table 4. Specifically, five annotators with different backgrounds, including researchers, master's students in affective computing, and general users, rated each sample from 1 to 5 based on emotional consistency and fluency. The final score is calculated as the average across annotators. In our experiments, REAR significantly improved the model's consistency while preserving its generalization ability, with no noticeable decline in empathy or fluency.



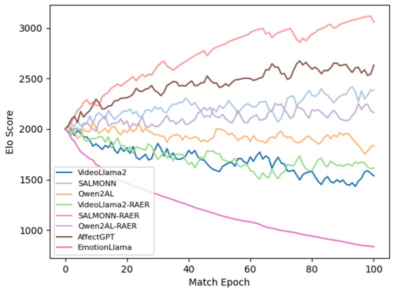
(a) Mixed Task on Visual-Language



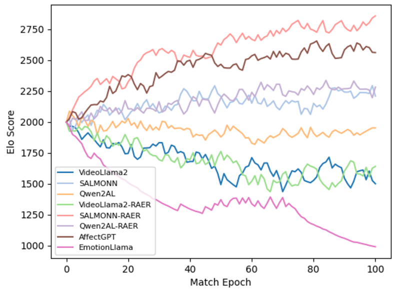
(b) MER Task on Visual-Language



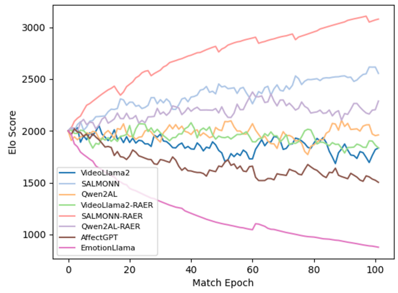
(c) MERG Task on Visual-Language



(d) Mixed Task on Audio-Language



(e) MER Task on Audio-Language



(f) MERG Task on Audio-Language

Figure 5: Part of the SAB evaluation results, where we used three task pools: Mixed, MER, and MERG. Figures (a), (b), and (c) evaluate visual-language capabilities, while figures (d), (e), and (f) assess audio-language capabilities. Most models' Elo scores remain relatively stable between upper and lower bounds, indicating that the models' abilities are accurately assessed through multiple rounds of testing.

	Emp./Con./Flu. (Human Evaluation)	
MLLMs	w/o RAER	RAER
AffectGPT	2.32/3.29/2.45	-
EmotionLlama	2.24/3.37/2.37	-
VideoLlama2	3.42/3.22/3.53	3.44/3.34/3.54
LlavaNextVideo	2.72/2.92/3.12	2.78/3.07/3.13
Qwen2.5VL	**3.76/3.46/3.78**	**3.74/3.78/3.76**
SALMONN	3.17/3.25/3.38	3.16/3.35/3.23
Qwen2Audio	3.24/3.32/3.42	3.26/3.42/3.35

Table 4: The results of the human evaluation of MERG tasks are presented, with all outcomes obtained through zero-shot inference.

5.3 Stimulus-Armed Bandit Evaluation

In the Stimulus-Armed Bandit (SAB) evaluation, we randomly generate one or a few emotion-neutral keywords and use generative models to produce corresponding stimuli. For this process, we employ GPT-4o to generate prompts, Sora for visual stimuli, and AudioGen for audio stimuli; further details can be found in Appendix C. Each evaluation match randomly selects a task from a predefined task pool, which is then combined with the generated stimuli to create a task prompt. The paired models are then tasked with providing the most suitable responses to the prompt. These responses are evaluated based on human or GPT-4o preferences, and the SAB framework uses an Elo-based scoring mechanism to dynamically update model scores.

As shown in Figure 5, We evaluate both MER and MERG tasks under two input conditions: visual-language (VL) and audio-language (AL). In the VL setting, RAER-enhanced models such as Qwen2.5-VL and VideoLLaMA2 demonstrate clear performance advantages over other models. Even without RAER, these models still show strong generalization capabilities, achieving results comparable to AffectGPT, which is explicitly fine-tuned for emotion tasks. This suggests that for well-generalized models, RAER can further improve emotion task performance by injecting relevant reasoning content. In the AL setting, RAER-enhanced Qwen2-Audio achieves the highest performance, while other models struggle to surpass AffectGPT. This suggests that due to limited generalization in the audio-language modality, current models—despite being enhanced with RAER—are still unable to outperform supervised fine-tuning approaches on emotion tasks, even when those fine-tuned models are trained with only limited data.

It is worth noting that Qwen2.5-VL, Qwen2-Audio, and VideoLLaMA2 are all based on Qwen2.5 or Qwen2 as the backbone LLMs, which have demonstrated strong generalization in prior studies. This may explain why RAER brings more significant gains when applied to Qwen-based mod-

Task Prompt:
The man is reading under the starry sky. Analyze the person's emotions.

Preferred:
The man in the image appears focused and engaged in reading, with his gaze directed upwards, indicating concentration. The setting in the background, with the starry sky, might evoke a sense of curiosity or contemplation, suggesting a calm and reflective emotional state.

Non-Preferred:
Although the man seems engaged in his reading, his slightly furrowed brow and neutral expression might indicate a sense of confusion or frustration, particularly if the reading material is challenging or not well-understood. The ambient lighting and background could imply a more tense or serious emotional state.

Figure 6: An example of SAB collected human preferences and corresponding task labels.

els. Although these MLLMs perform similarly on standard multimodal tasks, as emotional reasoning complexity increases, the base LLM's generalization ability becomes increasingly critical for downstream performance.

In our experiments, most SAB task pools converge within approximately 100 rounds, and the resulting model rankings are consistent with those from conventional benchmark evaluations. This demonstrates that the SAB framework effectively captures human-like preferences and serves as a comprehensive and scalable evaluation method for emotional reasoning and generation in MLLMs.

5.4 Compound Emotion QA Construction

Data type	Human Preference	GPT-4 Preference
VL	240	760
AL	1000	0

Table 5: Number of samples for Visual-Language (VL) and Audio-Language (AL) datasets, with preferences by human evaluators and GPT-4.

	MER2024	DFEW	IEMOCAP
VideoLlama2	0.35	0.32	0.41
w/ DPO	0.48	0.41	0.53
w/ RAER	0.62	**0.66**	0.71
w/ DPO&RAER	**0.67**	0.65	**0.79**

Table 6: The ablation experiment on DPO and RAER, with all metrics evaluated using WAR.

We compiled the results from the SAB evaluation into two sub-datasets, as shown in Table 5. Each sample is annotated with preferred and non-preferred responses, as illustrated in Figure 6. In the SAB framework, a generative model is used to create diverse multimodal scenarios, each requiring models to perform MER or MERG tasks. The target model conducts multimodal inference on the generated stimuli—videos, audio, and text—and generates corresponding responses. For each round, we then select the highest-ranked response as the *preferred* sample, and a median-quality response as the *non-preferred* counterpart, forming a pairwise preference instance. Based on this, we fine-tuned VideoLLama2 using Direct Preference Optimization (DPO) and conducted experiments on the MER task. The ablation experiment results are presented in Table 6. The findings suggest that our Compound Emotion QA dataset leads to an improvement in model performance.

6 Conclusion

In this paper, we introduced the Retrieval-Augmented Emotion Reasoning (RAER) framework, which enhances multimodal large language models (MLLMs) by combining emotional reasoning with retrieval-augmented processes. Our experiments on standard benchmarks and the Stimulus-Armed Bandit (SAB) evaluation demonstrate RAER's effectiveness in handling compound emotion tasks. We also contributed the Compound Emotion QA dataset, an AI generated dataset designed to further improve emotional reasoning. The results highlight RAER's potential in advancing multimodal sentiment analysis and enhancing human-AI interaction.

Limitations

Although our study presents promising advancements, it is not without its limitations. Firstly, while we aim for full automation, current state-of-the-art multimodal models still fall short in aligning with human preferences in audio. Although preference-based methods reduce the cost of manual selection, SAB cannot yet be fully automated. Secondly, RAER requires longer inference times compared to regular reasoning methods, leading to significantly lower computational efficiency. We are exploring more advanced CoT methods to potentially replace the current approach. Additionally, we utilize generative models to generate samples; however, there is currently no VAT-enabled (Visual,Audio,Text) generative model that covers all three modalities. As such, the SAB evaluation framework cannot fully replace traditional evaluation methods at this stage. Lastly, while RAER leverages human preferences at the annotation level, we are considering extending this by incorporating human feedback learning to further capitalize on the preference data generated during the RAER process.

Acknowledgments

This work is supported by the Strategic Priority Research Program of Chinese Academy of Sciences, Grant No. XDB0500103, the National Natural Science Foundation of China(NSFC) (No.62276259,No.62201572,No.62271083, No.U21B2010).

References

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional ai: Harmlessness from ai feedback. *Preprint*, arXiv:2212.08073.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. Narayanan. 2008. Iemocap: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359.

Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Jingdong Sun, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander Hauptmann. 2024a. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. *arXiv preprint arXiv:2406.11161*.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. 2024b. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & Emotion*, 6:169–200.

Hao Fei, Han Zhang, Bin Wang, Lizi Liao, Qian Liu, and Erik Cambria. 2024. EmpathyEar: An open-source avatar multimodal empathetic chatbot. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 61–71, Bangkok, Thailand. Association for Computational Linguistics.

Esam Ghaleb, Mirela Popa, and Stylianos Asteriadis. 2019. Multimodal and temporal perception of audio-visual cues for emotion recognition. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 552–558.

Yuan Gong, Yu-An Chung, and James Glass. 2021. Ast: Audio spectrogram transformer. *Preprint*, arXiv:2104.01778.

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*, MM '20, page 1122–1131, New York, NY, USA. Association for Computing Machinery.

Le Huang, Hengzhi Lan, Zijun Sun, Chuan Shi, and Ting Bai. 2024a. Emotional rag: Enhancing role-playing agents through emotional retrieval. *Preprint*, arXiv:2410.23041.

Yuming Huang, Bingfeng Ge, Zeqiang Hou, Hui Xie, Keith W. Hipel, and Kewei Yang. 2024b. Inverse preference optimization in the graph model for conflict resolution with uncertain cost. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 54(9):5580–5592.

Zhengjie Huang, Pingsheng Liu, Gerard de Melo, Liang He, and Linlin Wang. 2024c. Generating persona-aware empathetic responses with retrieval-augmented prompt learning. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12441–12445.

Zhilin Huang, Ling Yang, Xiangxin Zhou, Chujun Qin, Yijie Yu, Xiawu Zheng, Zikun Zhou, Wentao Zhang, Yu Wang, and Wenming Yang. 2024d.

Interaction-based retrieval-augmented diffusion models for protein-specific 3d molecule generation. In *International Conference on Machine Learning*.

Xingxun Jiang, Yuan Zong, Wenming Zheng, Chuangao Tang, Wanchuang Xia, Cheng Lu, and Jiateng Liu. 2020. Dfew: A large-scale database for recognizing dynamic facial expressions in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2881–2889.

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. 2022. Audiogen: Textually guided audio generation. *arXiv preprint arXiv:2209.15352*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Zheng Lian, Haiyang Sun, Licai Sun, Hao Gu, Zhuofan Wen, Siyuan Zhang, Shun Chen, Mingyu Xu, Ke Xu, Kang Chen, Lan Chen, Shan Liang, Ya Li, Jiangyan Yi, Bin Liu, and Jianhua Tao. 2024a. Explainable multimodal emotion recognition. *Preprint*, arXiv:2306.15401.

Zheng Lian, Haiyang Sun, Licai Sun, Zhuofan Wen, Siyuan Zhang, Shun Chen, Hao Gu, Jinming Zhao, Ziyang Ma, Xie Chen, Jiangyan Yi, Rui Liu, Kele Xu, Bin Liu, Erik Cambria, Guoying Zhao, Björn W. Schuller, and Jianhua Tao. 2024b. Mer 2024: Semi-supervised learning, noise robustness, and open-vocabulary multimodal emotion recognition. *Preprint*, arXiv:2404.17113.

Zheng Lian, Haiyang Sun, Licai Sun, Jiangyan Yi, Bin Liu, and Jianhua Tao. 2024c. Affectgpt: Dataset and framework for explainable multimodal emotion recognition. *Preprint*, arXiv:2407.07653.

Zheng Lian, Licai Sun, Mingyu Xu, Haiyang Sun, Ke Xu, Zhuofan Wen, Shun Chen, Bin Liu, and Jianhua Tao. 2023. Explainable multimodal emotion reasoning. *arXiv preprint arXiv:2306.15401*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Zhiwei Liu, Kailai Yang, Qianqian Xie, Christine de Kock, Sophia Ananiadou, and Eduard Hovy. 2024. Raemollm: Retrieval augmented llms for cross-domain misinformation detection using in-context learning based on emotional information. *Preprint*, arXiv:2406.11093.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Flor Miriam Plaza Del Arco, Amanda Curry, Alba Cercas Curry, Gavin Abercrombie, and Dirk Hovy. 2024. Angry men, sad women: Large language models reflect gendered stereotypes in emotion attribution. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7682–7696, Bangkok, Thailand. Association for Computational Linguistics.

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. *Preprint*, arXiv:2103.00020.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Preprint*, arXiv:2305.18290.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

James A Russell and Albert Mehrabian. 1977. Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11(3):273–294.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June Liu, Jinfeng Zhou, Alvionna Sunaryo, Tatia Lee, Rada Mihalcea, and Minlie Huang. 2024. EmoBench: Evaluating the emotional intelligence of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5986–6004, Bangkok, Thailand. Association for Computational Linguistics.

Junyoung Seo, Susung Hong, Wooseok Jang, Inès Hyeonsu Kim, Minseop Kwak, Doyup Lee, and Seungryong Kim. 2024. Retrieval-augmented score distillation for text-to-3d generation. *arXiv preprint arXiv:2402.02972*.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2022. Learning to summarize from human feedback. *Preprint*, arXiv:2009.01325.

Guangzhi Sun, Wenyi Yu, Changli Tang, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, Yuxuan

Wang, and Chao Zhang. 2024. video-SALMONN: Speech-enhanced audio-visual large language models. In *Forty-first International Conference on Machine Learning*.

Qwen Team. 2025. Qwen2.5-vl.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497.

Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Learning factorized multimodal representations. In *ICLR*.

Hanyang Wang, Bo Li, Shuang Wu, Siyuan Shen, Feng Liu, Shouhong Ding, and Aimin Zhou. 2023. Rethinking the learning paradigm for dynamic facial expression recognition. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17958–17968.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Jiaqiang Wu, Xuandong Huang, Zhouan Zhu, and Shangfei Wang. 2025. From traits to empathy: Personality-aware multimodal empathetic response generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8925–8938, Abu Dhabi, UAE. Association for Computational Linguistics.

Wen Wu, Bo Li, Chao Zhang, Chung-Cheng Chiu, Qiujia Li, Junwen Bai, Tara Sainath, and Phil Woodland. 2024. Handling ambiguity in emotion: From out-of-domain detection to distribution estimation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2078–2093, Bangkok, Thailand. Association for Computational Linguistics.

Zhou Yang, Zhaochun Ren, Wang Yufeng, Haizhou Sun, Chao Chen, Xiaofei Zhu, and Xiangwen Liao. 2024. An iterative associative memory model for empathetic response generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3081–3092, Bangkok, Thailand. Association for Computational Linguistics.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Memory fusion network for multi-view sequential learning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Yiqun Zhang, Fanheng Kong, Peidong Wang, Shuang Sun, SWangLing SWangLing, Shi Feng, Daling Wang, Yifei Zhang, and Kaisong Song. 2024a. STICKERCONV: Generating multimodal empathetic responses from scratch. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7707–7733, Bangkok, Thailand. Association for Computational Linguistics.

Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. 2024b. Llava-next: A strong zero-shot video understanding model.

Zhonghua Zheng, Lizi Liao, Yang Deng, Libo Qin, and Liqiang Nie. 2024. Self-chats from large language models make small emotional support chatbot better. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11325–11345, Bangkok, Thailand. Association for Computational Linguistics.

A Hyperparameters

Key hyperparameters used for RAER, SAB, and DPO are provided in Table 7..

B MER Experiments Details

Table 9 reports detailed evaluation metrics for each dataset in the MER task.

C SAB Task Formulation

As shown in Figure 7, we first use GPT-4 generates one or more emotionally neutral words. These neutral words are then used to generate corresponding modality-specific stimuli for the generative model. Afterward, the corresponding task and stimuli are matched, forming an SAB sample, the tasks are shown in Table 10 and Table 11.

As shown in Table 8, we tested GPT-4o's alignment with human preferences in the SAB visual-language evaluation. The experiment shows that GPT-4o is largely aligned with human preferences and can automatically evaluate and generate positive and negative samples.

D Why Not Fine-tune with SFT?

While it is theoretically possible to fine-tune an LLM for emotion tasks, we avoided this for two key reasons. First, supervised fine-tuning requires a significant amount of high-quality, open-ended,

Module	Hyperparameter	Value
RAER	K	5
SAB	K	64
	M	600
	new_output_token	500
	temperature	0.7
	top_p	0.9
DPO	seed	111
	torch_dtype	bfloat16
	max_length	1024
	tokenizer_name_or_path	`DAMO-NLP-SG/VideoLLaMA2.1-7B-16F-Base`
	loss_type	sigmoid
	beta	0.1
	per_device_train_batch_size	4
	gradient_accumulation_steps	4
	warmup_steps	100
	learning_rate	5e-6
	weight_decay	0.01
	num_train_epochs	3
	max_grad_norm	1.0
	lora_r	16
	lora_alpha	64
	lora_dropout	0.05
	target_modules	`{q_proj, k_proj, v_proj, o_proj, up_proj, down_proj, gate_proj}`

Table 7: Hyperparameter Settings for RAER, SAB, and DPO

Alignment	Number of Samples	Rate
Consistent	228	76%
Inconsistent	72	24%

Table 8: GPT-4o's alignment with human preferences in SAB visual-language evaluation.

human-annotated emotional data. However, current datasets such as EMER and MERR contain only limited samples (e.g., 332 annotated responses in EMER), which are insufficient for stable and effective fine-tuning. Second, in our prior studies, we observed that SFT on current emotion datasets (e.g. EMER,MERR) tends to lead to overfitting within the limited training domain, thereby impairing generalization ability and degrading the performance of chain-of-thought (CoT) reasoning. Furthermore, RAER's retrieval-augmented design is more effective when the model retains its general-purpose inference capacity, which can be compromised by fine-tuning on small or narrow datasets. Thus, we choose to apply RAER in a zero-shot setting across general-purpose MLLMs.

	MER2024 w/o RAER								
	Acc↑	Acc↑	Acc↑	Acc↑	Acc↑	Acc↑	Acc↑	UAR↑	WAR↑
MLLMs	Worried	Happy	Neutral	Angry	Surprise	Sad			
AffectGPT	0.87	0.83	0.00	0.67	0.12	0.78		0.57	0.64
EmotionLlama	0.55	0.12	0.00	0.25	0.06	0.19		0.21	0.23
VideoLlama2	0.76	0.38	0.26	0.21	0.00	0.40		0.34	0.35
LlavaNextVideo	0.62	0.32	0.21	0.12	0.00	0.08		0.22	0.23
Qwen2.5VL	0.20	0.58	1.00	0.06	0.00	0.38		0.37	0.45
SALMONN	0.52	0.48	0.32	0.15	0.00	0.12		0.28	0.31
Qwen2Audio	0.21	0.39	0.62	0.12	0.00	0.11		0.25	0.27
	MER2024 RAER								
MLLMs	Worried	Happy	Neutral	Angry	Surprise	Sad			
VideoLlama2	0.83	0.68	0.83	0.42	0.00	0.54		0.56	0.62
LlavaNextVideo	0.74	0.33	0.42	0.24	0.00	0.08		0.33	0.37
Qwen2.5VL	0.78	0.84	0.92	0.66	0.00	0.46		0.65	0.73
SALMONN	0.56	0.48	0.32	0.27	0.00	0.35		0.39	0.42
Qwen2Audio	0.38	0.45	0.72	0.16	0.00	0.09		0.33	0.35
	DFEW w/o RAER								
MLLMs	Happy	Fear	Neutral	Angry	Surprise	Sad	Disgust		
AffectGPT	0.83	0.64	0.00	0.52	0.21	0.69	0.00	0.41	0.52
EmotionLlama	0.15	0.12	0.00	0.25	0.09	0.19	0.00	0.14	0.59
VideoLlama2	0.34	0.28	0.26	0.21	0.22	0.25	0.00	0.25	0.32
LlavaNextVideo	0.18	0.12	0.11	0.12	0.07	0.08	0.00	0.14	0.22
Qwen2.5VL	0.59	0.38	0.85	0.24	0.32	0.31	0.00	0.38	0.43
SALMONN	0.25	0.12	0.21	0.23	0.11	0.07	0.00	0.17	0.24
Qwen2Audio	0.32	0.16	0.24	0.07	0.02	0.11	0.00	0.14	0.22
	DFEW RAER								
MLLMs	Happy	Fear	Neutral	Angry	Surprise	Sad	Disgust		
VideoLlama2	0.63	0.44	0.57	0.52	0.47	0.39	0.00	0.45	0.66
LlavaNextVideo	0.32	0.19	0.25	0.17	0.25	0.22	0.00	0.24	0.32
Qwen2.5VL	0.64	0.49	0.92	0.36	0.37	0.48	0.00	0.49	0.69
SALMONN	0.31	0.17	0.32	0.29	0.21	0.17	0.00	0.26	0.37
Qwen2Audio	0.33	0.19	0.23	0.14	0.03	0.19	0.00	0.22	0.28
	IEMOCAP w/o RAER								
MLLMs	Happy	Sad	Neutral	Angry					
AffectGPT	0.90	0.84	0.00	0.77				0.62	0.71
EmotionLlama	0.25	0.43	0.00	0.25				0.25	0.32
VideoLlama2	0.42	0.38	0.43	0.22				0.38	0.41
LlavaNextVideo	0.33	0.22	0.30	0.14				0.27	0.28
Qwen2.5VL	0.51	0.42	0.86	0.31				0.51	0.53
SALMONN	0.35	0.31	0.22	0.26				0.31	0.34
Qwen2Audio	0.45	0.22	0.34	0.15				0.30	0.31
	IEMOCAP RAER								
MLLMs	Happy	Sad	Neutral	Angry					
VideoLlama2	0.77	0.68	0.69	0.57				0.70	0.71
LlavaNextVideo	0.51	0.32	0.44	0.28				0.37	0.40
Qwen2.5VL	0.85	0.72	0.85	0.61				0.77	0.78
SALMONN	0.44	0.36	0.23	0.35				0.39	0.39
Qwen2Audio	0.57	0.25	0.39	0.17				0.36	0.37

Table 9: Detailed results of MER tasks are presented. All results are obtained using zero-shot inference.

Task	Description
Multi-choice	Select the most appropriate emotions from a predefined list based on multimodal inputs.
Ranking	Rank emotions from a predefined list based on multimodal inputs.
Recognition Analysis	Analyze and identify emotions in the multimodal input.
Transition Detection	Detect emotional shifts over time and identify when and how emotions change in a sequence.

Table 10: Multimodal Emotion Recognition (MER) Task Set Description

Task	Description
Emotion-based Response Generation	How would you feel if you were in the person's shoes in this video?
Emotion-based Response Generation	How does the person's experience in the video make you feel?
Emotion-based Response Generation	How does the video make you reflect on your own emotions or experiences?
Emotion-based Response Generation	What would you be feeling right now?

Table 11: Multimodal Empathetic Response Generation (MERG) Task Set Description

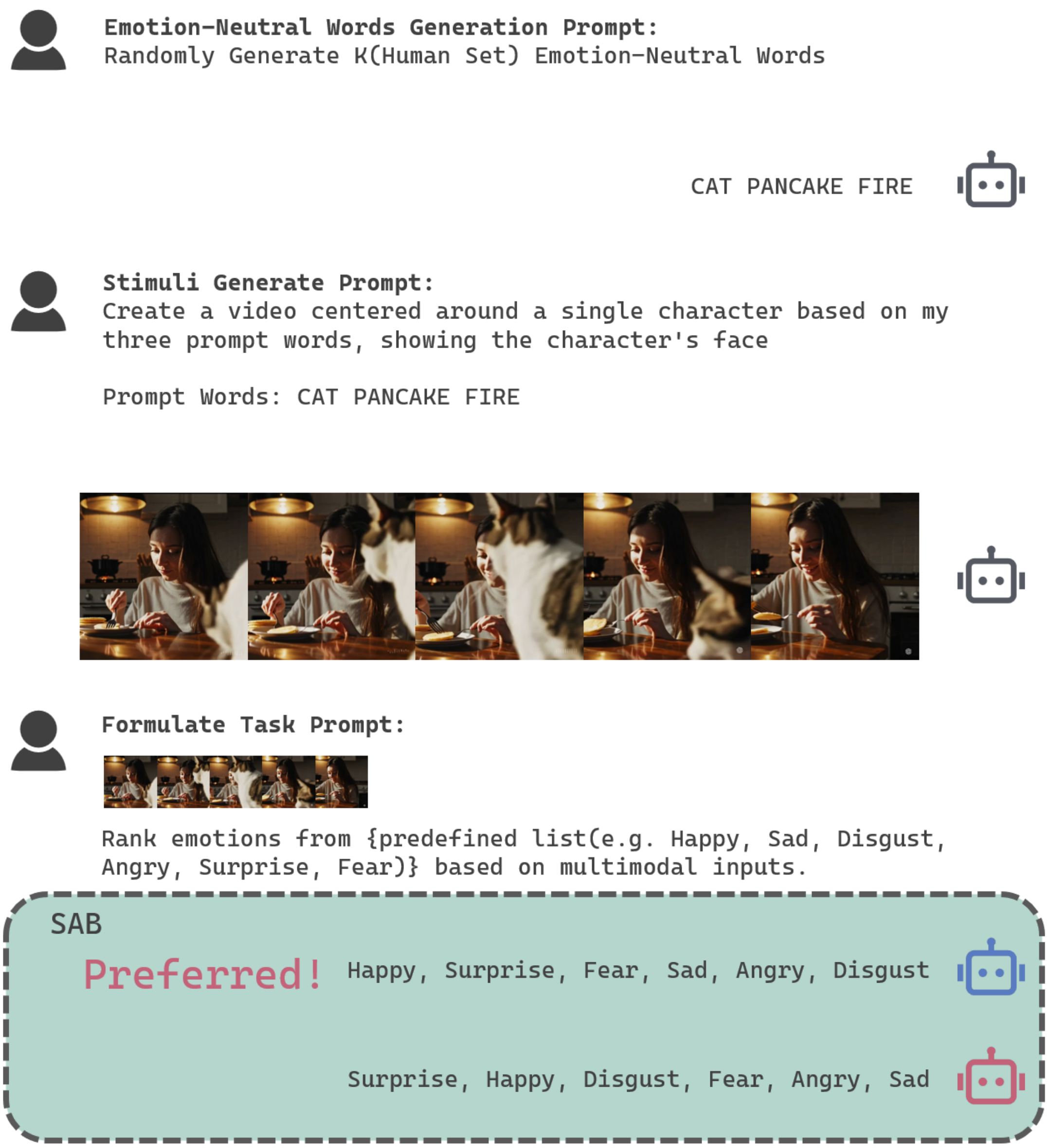


Figure 7: An example of SAB task formulation