

# Emotion-Coherent Reasoning for Multimodal LLMs via Emotional Rationale Verifier

Hyeongseop Rha, Jeong Hun Yeo, Yeonju Kim, Yong Man Ro\*

Integrated Vision and Language Lab, KAIST, South Korea  
{ryool\_1832, sedne246, yeonju7.kim, ymro}@kaist.ac.kr

## Abstract

The recent advancement of Multimodal Large Language Models (MLLMs) is transforming human-computer interaction (HCI) from surface-level exchanges into more nuanced and emotionally intelligent communication. To realize this shift, emotion understanding becomes essential allowing systems to capture subtle cues underlying user intent. Furthermore, providing faithful explanations for predicted emotions is crucial to ensure interpretability and build user trust. However, current MLLM-based methods often generate emotion explanations that diverge from the target labels and sometimes even contradict their own predicted emotions. This inconsistency poses a critical risk for misunderstanding and erodes reliability in interactive settings. To address this, we propose a novel approach: the Emotional Rationale Verifier (ERV) and an Explanation Reward. Our method guides the model to produce reasoning that is explicitly consistent with the target emotion during multimodal emotion recognition without modifying the model architecture or requiring additional paired video–description annotations. Our method significantly improves faithful explanation–prediction consistency and explanation emotion accuracy on the MAFW and DFEW datasets. Through extensive experiments and human evaluations, we show that our approach not only enhances alignment between explanation and prediction but also empowers MLLMs to deliver emotionally coherent, trustworthy interactions, marking a key step toward truly human-like HCI systems.

**Code** — <https://github.com/Rhatanii/ERV>

**Extended version** — <https://arxiv.org/abs/2510.23506>

## 1 Introduction

As multimodal large language models (MLLMs) (Chen et al. 2024; Li et al. 2024; Wang et al. 2024; Chu et al. 2024; Cheng et al. 2024b) continue to evolve, they are reshaping human-computer interaction (HCI) by enabling systems to interpret cross-modal inputs (face video and speech audio), perform contextual reasoning, and generate human-like responses (Fei et al. 2024; Park et al. 2024; Kim et al. 2025). To support truly natural and emotionally aware communication, such systems must go beyond surface-level responses and

capture the nuance, intent, and emotional context of user inputs. Among these capabilities, emotion understanding plays a central role in facilitating deeper, more meaningful interaction—particularly in applications such as psychological counseling (Schmidgall et al. 2024; Lee et al. 2024), educational guidance (Schutz and Pekrun 2007), and empathetic dialogue systems (Fei et al. 2024; Zhang et al. 2025).

While conventional multimodal emotion recognition (MER) approaches have treated emotion understanding as a classification task (Zhang et al. 2024b), single-label predictions often fail to capture the complexity of human affect conveyed through face video and speech audio. To address this limitation, recent work has turned to *descriptive emotion understanding*, leveraging MLLMs to generate natural language explanations that justify predicted emotions (Lian et al. 2023; Cheng et al. 2024a; Lian et al. 2025; Yang et al. 2025a). This shift is especially valuable in HCI scenarios, where users not only expect accurate emotional classification from visual and auditory cues but also seek coherent and human-like justifications to build trust and ensure transparency.

To enhance emotion reasoning, a recent study leverages Reinforcement Learning (RL) and Chain-of-Thought (CoT) prompting to train MLLMs (Zhao, Wei, and Bo 2025). Although this method improves classification accuracy, it fails to consistently produce emotionally coherent explanations. This limitation may be attributable to the datasets used in current training pipelines. While recent method leverages a variety of emotion datasets, such as EMER (Lian et al. 2023), MERR (Cheng et al. 2024a), and MER-Caption (Lian et al. 2025), only a small fraction of these include human-annotated rationales. The majority of training relies on larger datasets without explanation labels, where reward signals focus solely on the correctness of predicted emotions, regardless of whether the generated rationale is emotionally coherent. As a result, models generate emotionally incoherent explanations without penalty, leading to a misalignment between predictions and rationales, as shown in Figure 1.

In this paper, we propose a novel framework for generating emotionally coherent explanations, even when trained on emotion datasets lacking human-annotated rationales. Our approach is based on the key assumption that LLMs possess sufficient linguistic and emotional knowledge to evaluate the consistency between a given emotion and a generated explanation. However, incorporating LLMs directly into the

\*Corresponding author.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

**Prompt:** As an emotional recognition expert; throughout the video, which emotion conveyed by the characters is the most obvious to you? Output the thinking process in `<think>` `</think>` and final emotion in `<answer>` `</answer>` tags.

“Yes. The most intense eyes. He doesn’t move and he is just like a thunder.”

Input Video



Baseline (R1-Omni)

`<think>` In the video, we see a woman . . . facial expression appears somewhat **surprised or confused**. In the audio, it is difficult to determine the emotional state of the person. In the text, the subtitle says, “He has the most intense eyes.” This sentence may be the woman’s evaluation or reaction to someone . . . Based on the woman’s facial expressions in the video clues, which display **surprise or confusion** . . . may carry a **negative** emotion. `</think>` `<answer>` **happy** `</answer>`

✗ Explanation: **Surprise**      ✓ Answer: **Happy**

With ERV (Ours)

`<think>` In the video, we see a woman . . . slight **smile** on her face, as if she is speaking or thinking. Her eyes are wide open, seemingly emphasizing something . . . seems like she is engaged in a **pleasant** conversation. In the audio, the character’s tone is gentle and the volume is high, suggesting that the character is **happy** and **satisfied**. In the text, the subtitle says, “He has the most intense eyes.” . . . possibly expresses **praise** or **admiration** for the other person’s eyes . . . this sentence may express the woman’s **appreciation** and praise for the other person’s eyes. `</think>` `<answer>` **happy** `</answer>`

✓ Explanation: **Happy**      ✓ Answer: **Happy**

Figure 1: Illustrative comparison between the baseline and our model on explanation–emotion alignment (text truncated for brevity). While both correctly predict the final emotion, the baseline fails to generate an emotionally coherent explanation, highlighting its misalignment.

evaluation process can be computationally expensive and impractical in deployment settings. To address this, we distill the knowledge of the LLM into a lightweight model, referred to as the Emotional Rationale Verifier (ERV), which efficiently assesses the alignment between an explanation and the target emotion. Building on this verifier, we integrate ERV into a reward-driven training pipeline. In this setup, the model receives an explanation reward based on how well its generated explanation aligns with the target emotion as judged by ERV. To further improve reward fidelity, we decompose each explanation into individual sentences, filter out emotionally neutral ones (e.g., those describing background or appearance), and assign rewards based on the proportion of emotionally salient sentences.

Furthermore, to evaluate the emotional coherence, we introduce a set of novel metrics: Explanation Emotion Accuracy (EEA), which measures alignment between the explanation and the target emotion, Faithful Consistency Rate (FCR), requires all three components, explanation, predicted emotion, and target emotion, to agree, and Explanation–Prediction Consistency (EPC), which quantifies the consistency between the explanation and predicted emotion. These metrics provide deeper insight into the emotional coherence of generated explanations beyond standard classification accuracy.

Experimental results demonstrate that our approach preserves emotion classification performance while improving the consistency between generated explanations and predicted emotions, as measured by the three proposed evaluation metrics. Moreover, human evaluation shows that the model produces more coherent and emotionally aligned reasoning, thereby supporting more reliable and interpretable interactions in HCI applications.

Our key contributions are as follows:

- **Emotionally consistent explanation framework:** We propose a novel framework that improves the emotional coherence of explanations without modifying the model architecture or requiring labeled explanation data by leveraging CoT reasoning and reward-based learning.
- **Emotional Rationale Verifier (ERV):** We introduce ERV, a lightweight verifier distilled from an LLM, which assesses the alignment between generated explanations and target emotions, and serves as a reward function to guide explanation generation.
- **New metrics for emotional coherence evaluation:** We propose three novel metrics—EEA, EPC, and FCR—that enable fine-grained evaluation of emotional alignment and consistency in model-generated explanations.

## 2 Related Works

### 2.1 Multimodal Emotion Recognition

The task of Multimodal Emotion Recognition (MER) aims to understand human emotional states by analyzing diverse multimodal cues in expressive scenes. From audio and visual cues to text, body, and physiological signals, numerous studies (Kossaifi et al. 2017; Zadeh et al. 2018; Poria et al. 2019; Busso et al. 2008; Perepelkina, Kazimirova, and Konstantinova 2018; Liu et al. 2022; Jiang et al. 2020; Livingstone and Russo 2018; Lee et al. 2019; Wang et al. 2022) have proposed datasets consisting of emotion-related modalities and corresponding emotion labels. With those datasets, traditional approaches have primarily relied on modality-specific encoders (Zhang et al. 2017, 2021; Zhao and Liu 2021; Sun et al. 2023; Chumachenko, Iosifidis, and Gabbouj 2024) for

feature extraction and fusion strategies (Narayan et al. 2025; Sun et al. 2024), achieving solid performance in emotion classification.

However, unlike object classification, emotion classification suffers from inherent ambiguity, as emotion labels alone often fail to fully capture an individual’s state. To address the challenge of unreliable labeling caused by the limited expressivity of fixed label mappings, multimodal emotion descriptive dataset EMER (Lian et al. 2023) has been proposed. This dataset requires models not only to predict emotion labels from visual, audio, and textual cues, but also to generate descriptive explanations of the emotional states. Additionally, with the rise of Large Language Models (LLMs) (Touvron et al. 2023; Team 2024; Grattafiori et al. 2024), and their powerful natural language generation capabilities, recent instruction-tuned MLLM approaches (Xing et al. 2024; Cheng et al. 2024a; Lian et al. 2025; Yang et al. 2025a; Hu et al. 2025; Zhao et al. 2025a) move beyond classification toward reasoning, enabling interpretable and context-aware emotion understanding.

Furthermore, motivated by the increasing use of rich emotional descriptions, recent studies (Cheng et al. 2024a; Lian et al. 2024; Yang et al. 2025a) have explored leveraging the generative capabilities of LLMs to automatically expand emotion descriptive datasets. These approaches aim to enrich the diversity and expressivity of descriptive samples by generating additional explanations grounded in multimodal cues. Building on such augmented and description-rich emotion datasets, many works (Cheng et al. 2024a; Lian et al. 2024) have investigated interpretable reasoning over video and audio, while architectural advancements (Zhao et al. 2025b; Lian et al. 2025) have further improved recognition accuracy. However, most existing methods rely on supervised fine-tuning (SFT), which limits generalization due to its dependence on the scale and diversity of annotated emotional descriptions.

## 2.2 GRPO-Based RL with Verifiable Rewards

Recent work has shown that RL post-training can substantially sharpen the CoT abilities of MLLMs. In particular, Group Relative Policy Optimization (GRPO) (Shao et al. 2024) combined with Verifiable Reward (Guo et al. 2025) achieves strong generalization with limited data and fixed model backbones, excelling in math and coding benchmarks.

Motivated by this success, GRPO has been adapted to a wide range of multimodal tasks, including emotion recognition (Zhao, Wei, and Bo 2025), general vision tasks (Chen et al. 2025a), object detection (Liu et al. 2025; Shen et al. 2025; Park et al. 2025), and the video domain (Li et al. 2025; Lee et al. 2025; Xing et al. 2025), often designing reward functions tailored to their specific task objectives. However, these approaches predominantly focus on optimizing the final answer, paying little attention to the quality and consistency of the intermediate reasoning steps, a degradation problem highlighted by (Wei et al. 2025).

To address this, recent studies (Chen et al. 2025b; Yang et al. 2025b) have introduced consistency-aware methodologies to improve the alignment between reasoning and the answer. Yet, those consistency problems have not yet been

explored in the field of MER, where reasoning datasets are limited. This gap is significant, as MER applications require explanations that are not just plausible but emotionally coherent with the predicted affect. We therefore introduce a novel reward applied with GRPO training that explicitly enforces coherence between the generated explanation and the target emotion, addressing the unique challenge of achieving consistency in the absence of human-annotated rationales.

## 3 Method

Our primary objective is to develop an MLLM capable of emotion understanding that not only accurately predicts emotion labels from face video and speech audio but also generates consistent and faithful natural language explanations aligned with these predictions. Building upon recent advancements demonstrating the effectiveness of RL-based training for improving both emotion recognition accuracy and reasoning in MLLM (Zhao, Wei, and Bo 2025), we integrate a novel Emotional Rationale Verifier (ERV) module into an established RL framework. This integration allows us to directly evaluate and enhance the emotional consistency of the generated explanations. Specifically, we extend the HumanOmni model architecture (Zhao et al. 2025b) and GRPO (Guo et al. 2025), a reward-driven RL training strategy showing strong performance in various multimodal tasks (Liu et al. 2025; Chen et al. 2025a; Zhao, Wei, and Bo 2025). While HumanOmni effectively unifies multimodal features and GRPO enhances CoT reasoning, existing approaches often struggle with emotionally incongruent explanations despite accurate label predictions, as highlighted in our introduction.

To precisely address this critical misalignment, our proposed ERV module assesses whether a generated explanation coherently supports the target emotion. Based on this assessment, we define a novel reward signal, the Explanation Reward ( $R_E$ ), which encourages the MLLM to produce more faithful and emotion-aligned rationales. The overall training pipeline, detailing the ERV module and its integration into the GRPO framework, is visually represented in Figure 2.

### 3.1 Model Architecture and Training Overview

Our methodology leverages the HumanOmni architecture as its foundational MLLM. Given an input video  $V$ , audio  $A$ , and task prompt  $P$ , the model first processes the raw multimodal inputs using pretrained visual and audio encoders. These modality-specific features  $X_v$  (visual) and  $X_a$  (audio) are then projected into a shared embedding space and concatenated with the prompt embedding ( $X_p$ ). The resulting comprehensive fused representation ( $X_m$ ) is then fed into the Large Language Model (LLM), denoted as  $\psi$ , which autoregressively generates the output sequence ( $o$ ):

$$X_m = \text{Concat}(X_v, X_a, X_p), \quad o = \psi(X_m) \quad (1)$$

The training of this model follows a robust two-stage strategy: an initial Supervised Fine-Tuning (SFT) (Zhang et al. 2024a) phase, followed by RL using GRPO. This progressive approach ensures the model first learns fundamental reasoning structures and then refines its explanation generation based on explicit reward signals.

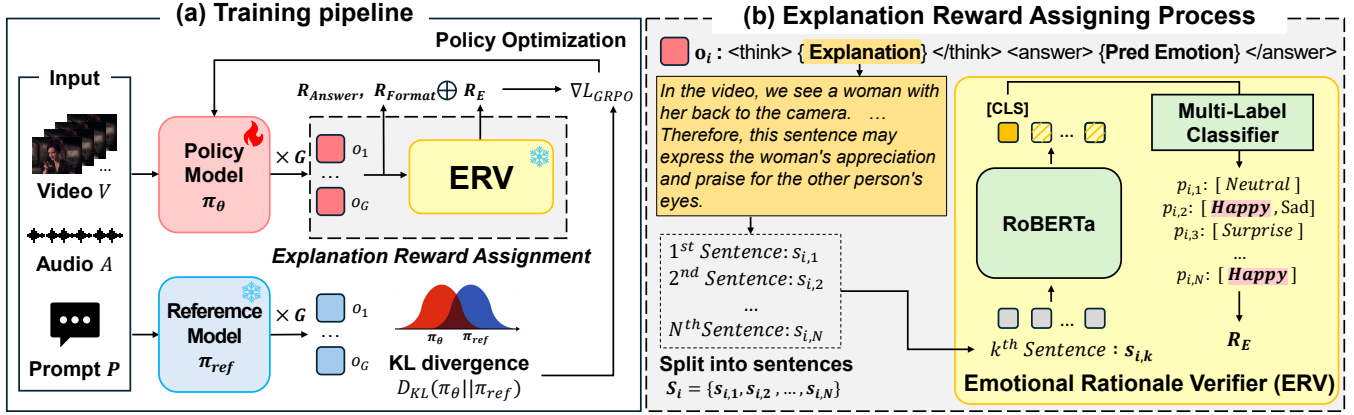


Figure 2: **(a)**: The policy model  $\pi_\theta$  generates  $G$  responses  $o_1, \dots, o_G$ , and the Emotional Rationale Verifier (ERV) assigns an explanation reward  $R_E$  to each response. **(b)**: For each response  $o_i$ , its explanation  $E_i$  is extracted and evaluated to produce its reward  $R_{i,E}$ . The ground truth emotion used for evaluation in this scenario is ‘Happy’ ( $e_{gt} = \text{‘Happy’}$ ). Collectively,  $R_E = \{R_{1,E}, R_{2,E}, \dots, R_{G,E}\}$  denotes the set of explanation rewards for the  $G$  responses.

**Initializing Emotion Reasoning via SFT** The initial training phase involves SFT on a high-quality, albeit relatively small, dataset named EMER (Lian et al. 2023). This dataset consists of  $(V, A)$  pairs annotated with emotion labels and detailed explanatory descriptions. The primary objective of this phase is to enable the MLLM to grasp the basic structural relationships between multimodal inputs, emotion labels, and their textual explanations. It specifically helps the model learn to produce outputs in the required format  $\langle \text{think} \rangle \dots \langle \text{/think} \rangle \langle \text{answer} \rangle \dots \langle \text{/answer} \rangle$ . This SFT phase serves as a crucial initialization step for our policy model, denoted as  $\pi_\theta$  (i.e., the trainable MLLM). During SFT, the model is optimized to maximize the likelihood of generating the target output sequence given the input. The training objective is the negative log-likelihood loss:

$$\mathcal{L}_{\text{SFT}} = -\mathbb{E}_{(x,y) \sim \mathcal{D}_{\text{SFT}}} \left[ \sum_{t=1}^{|y|} \log \pi_\theta(y_t | x, y_{<t}) \right] \quad (2)$$

Here,  $(x, y)$  represents the input-output pair sampled from the SFT dataset  $\mathcal{D}_{\text{SFT}}$ , where  $x$  represents the input triplet  $(V, A, P)$  and  $y$  is the corresponding explanation.

**Enhancing Explanation Coherence through Explanation-Guided Reinforcement Learning** Following the SFT phase, we transition to RL using GRPO to enhance the model’s reasoning capability and, critically, the consistency of its explanations. The SFT-trained model  $\pi_\theta$  functions as the learnable policy, while a frozen duplicate,  $\pi_{\text{ref}}$ , acts as the reference model. For each input  $(V, A, P)$ , both the policy model and the reference model generate  $G$  candidate output sequences,  $\{o_1, o_2, \dots, o_G\}$ . Each generated sequence is then evaluated by a comprehensive reward function, which we detail in the next section. The policy  $\pi_\theta$  is subsequently optimized to maximize the expected reward  $R$ , while simultaneously being regularized to remain close to the behavior of  $\pi_{\text{ref}}$ . The optimization objective of GRPO is as follows:

$$\max_{\pi_\theta} \mathbb{E}_{o \sim \pi_\theta(x)} [R(o) - \beta \cdot \text{KL}(\pi_\theta(o|x) || \pi_{\text{ref}}(o|x))] \quad (3)$$

While the GRPO framework is highly effective for reward-driven optimization, it traditionally lacks an explicit mechanism to guarantee the consistency between the predicted emotion and its generated explanation. To overcome this inherent limitation and directly address the core challenge identified in our introduction, we introduce the Emotional Rationale Verifier (ERV) and its associated reward ( $R_E$ ) as a novel and crucial signal within the GRPO training loop.

### 3.2 Emotional Rationale Verifier (ERV): Design and Training

We define an explanation-related reward signal based on how well the explanation in the output sequence  $o$  reflects the target emotion. To assess the emotional content of the generated explanation  $E$ , we employ an auxiliary module capable of classifying emotion labels from textual descriptions.

Although existing datasets such as EMER (Lian et al. 2023) and MERR-fine (Cheng et al. 2024a) provide pairs of emotional descriptions and emotion labels, their limited sizes (332 and 4,487 samples, respectively) are insufficient for our setting. This is particularly problematic because the generated descriptions often contain not only emotional expressions but also contextual details about people, environments, and temporally evolving emotional states. In addition, the emotion label distribution is imbalanced. As a result, existing datasets lack the diversity and coverage necessary to effectively train the ERV.

To address the limitations of existing datasets in terms of both size and class imbalance, we construct a pseudo-labeled dataset consisting of emotional text descriptions paired with emotion labels. As shown in Figure 1, while R1-Omni (Zhao, Wei, and Bo 2025) does not always produce perfectly emotion-aligned outputs, its generated descriptions are often emotionally plausible. Leveraging this property,

we generate textual explanations from the training sets of DFEW (Jiang et al. 2020) and MAFW (Liu et al. 2022), and use a closed-source LLM (GPT-4.1) to assign up to two representative emotion labels to each description. This process yields a pseudo-labeled dataset of 20K emotional text-label pairs, which is used to train our ERV module. To further mitigate the class imbalance, we augment underrepresented categories, namely *disgust*, *fear*, *contempt*, *disappointment*, *neutral*, and *helplessness*, by generating additional emotional descriptions using GPT-4.1.

To train the ERV on the constructed dataset, we use the RoBERTa model trained on the GoEmotions dataset (Demszky et al. 2020)<sup>1</sup> as a backbone. We fine-tune this model to predict emotion labels from text descriptions using our 23K-sample dataset, which includes both human-annotated and GPT-generated pseudo-labeled examples. The model predicts the emotion label based on the output of the [CLS] token, which encodes the global semantics of the input description. These predictions are then used to compute the Explanation Reward for each generated explanation  $E_i$ .

### 3.3 Explanation Reward

With the proposed ERV, we define an explanation reward  $R_E$ , which is assigned in proportion to how well the generated explanation  $E$  reflects the GT emotion  $e_{gt}$ . However, a long narrative often mixes various types of content such as scene background, appearance details, and affective cues, making it unreliable to evaluate the entire paragraph as a whole. Sentences that merely describe appearance or context can dilute the emotional signal and mislead ERV.

To address this, we design the explanation reward  $R_E$  with two key components. We illustrate the reward computation using the  $i$ th output  $o_i$  from the policy model  $\pi_\theta$ , as shown in Figure 2 (b). **1) Sentence-level and multi-label verification.** For each generated sentence  $s_{i,k} \in E_i$ , ERV predicts a (possibly multi-label) emotion set  $p_{i,k}$ . This sentence-level prediction prevents non-emotive sentences from dominating the overall judgment. Furthermore, since a single sentence may convey multiple emotions, we apply a threshold  $\tau$  to extract up to two emotion labels per sentence from the classifier. **2) Neutral-sentence filtering.** If ERV classifies  $s_{i,k}$  as NEUTRAL, the sentence is excluded from the computation of  $R_{i,E}$ , based on the intuition that surface-level descriptions such as listing physical traits or environmental details provide little to no emotional signal and thus do not meaningfully contribute to emotion understanding.

Given the set of split sentences  $S_i$ , each sentence is evaluated by ERV. We count the number of sentences whose predicted emotions  $p_{i,k}$  match the GT emotion  $e_{gt}$  as  $c_i$ , and denote the number of sentences classified as NEUTRAL as  $N_{i,Neu}$ . The explanation reward  $R_{i,E}$  is then computed as:

$$R_{i,E} = \begin{cases} \frac{c_i}{N_i - N_{i,Neu}}, & \text{if } e_{gt} \neq \text{NEUTRAL}, \\ \frac{c_i}{N_i}, & \text{if } e_{gt} = \text{NEUTRAL}. \end{cases} \quad (4)$$

**Total reward.** In addition to the  $R_E$ , we incorporate two auxiliary rewards *format reward*  $R_{format}$  and *answer reward*

$R_{answer}$  to compute the total reward  $R$ .  $R_{format}$ , following prior work (Guo et al. 2025), assigns 1 if the model output adheres to the required format (i.e., includes both `<think>` and `<answer>` tags), and 0 otherwise.  $R_{answer}$  assigns 1 if the predicted emotion within the `<answer>` matches  $e_{gt}$ , and 0 otherwise. The total reward used in training is computed as:

$$R = R_E + R_{format} + R_{answer}. \quad (5)$$

Incorporating  $R_E$  during GRPO training significantly enhances consistency between the model’s reasoning and the target emotion.

## 4 Experimental Setup

### 4.1 Dataset

**EMER** (Lian et al. 2023) is a human-annotated emotional video dataset with 332 samples labeled with one of five emotions (*angry*, *sad*, *surprise*, *worried*, *happy*) and corresponding descriptions. It is used to train the model to generate explanations in a reasoning format.

**MERR** (Cheng et al. 2024a) provides 4,487 fine-grained pseudo-labeled video-caption pairs, generated via facial and audio analysis with LLM-guided annotation. We use this data to supervise our ERV, leveraging its diverse and component-level emotion cues.

**DFEW** (Jiang et al. 2020) is a large-scale facial expression video dataset with 9,362 training and 2,342 test samples, labeled with 7 basic emotions: *angry*, *disgust*, *surprise*, *happy*, *sad*, *neutral*, and *fear* (official 5th split). We use DFEW in two ways: (1) generating textual explanations from training videos and assigning pseudo-labels via GPT-4.1 to augment ERV training, and (2) applying our reward-based RL scheme. **MAFW** (Liu et al. 2022) is a multimodal emotion dataset with 7 basic and 4 compound categories (*contempt*, *helplessness*, *anxiety*, *disappointment*), comprising 7,341 training and 1,831 test samples (set 5 split). We apply the same pipeline as DFEW.

### 4.2 Evaluation Metrics

**Emotion Recognition Accuracy** Following prior works (Zhao, Wei, and Bo 2025), we employ Unweighted Average Recall (UAR) and Weighted Average Recall (WAR) to measure the model’s emotion classification performance.

**Evaluating Emotional Coherence** We quantitatively evaluate the emotional coherence of generated explanations using three metrics: Explanation Emotion Accuracy (EEA), Explanation–Prediction Consistency (EPC), and Faithful Consistency Rate (FCR). These metrics capture different aspects of alignment among the explanation, predicted emotion, and ground-truth label. GPT-4.1-mini is used to recognize emotion in explanation. Figure 3 shows the detailed prompt used in evaluation.

### 4.3 Implementation Details

We employ the HumanOmni 0.5B model (Zhao et al. 2025b) as our backbone and RoBERTa-100M (trained on GoEmotions) for the ERV. Our training pipeline consists of three components: SFT, ERV, and GRPO. The main model is first fine-tuned on the EMER dataset for 5 epochs with a  $2e^{-5}$

<sup>1</sup>[https://huggingface.co/SamLowe/roberta-base-go\\_emotions](https://huggingface.co/SamLowe/roberta-base-go_emotions)

Model	MAFW					DFEW				
	EEA (%)	FCR (%)	EPC (%)	WAR (%)	UAR (%)	EEA (%)	FCR (%)	EPC (%)	WAR (%)	UAR (%)
<b>SFT-based Emotion MLLMs</b>										
Human-Omni (Zhao et al. 2025b)	-	-	-	20.18	13.52	-	-	-	22.64	19.44
EMER-SFT	33.06	29.67	<b>70.66</b>	38.95	25.21	35.14	32.63	64.30	38.93	30.86
EMER-SFT (7B)	44.26	38.58	67.38	50.22	33.48	44.86	42.21	57.41	48.33	39.16
Emotion-LLaMA* (7B) (Cheng et al. 2024a)	-	-	-	-	-	41.15	32.56	57.69	48.71	38.54
Emotion-LLaMA (7B) Fine-Tuned	-	-	-	-	-	-	-	-	77.06	64.21
<b>RL-based Emotion MLLMs</b>										
R1-Omni (Zhao, Wei, and Bo 2025)	37.87	31.80	48.09	<b>57.68</b>	<b>40.04</b>	43.69	37.99	51.91	65.83	<b>56.27</b>
Baseline	32.41	27.27	40.98	56.83	38.66	41.66	36.39	48.67	66.17	55.33
<b>Ours</b>	<b>48.93</b>	<b>43.57</b>	68.58	57.49	39.12	<b>59.45</b>	<b>53.75</b>	<b>73.06</b>	<b>67.30</b>	56.26

Table 1: Evaluation of models on proposed metrics: Explanation Emotion Accuracy (EEA), Faithful Consistency Rate (FCR), Explanation–Prediction Consistency (EPC), and classification accuracy (WAR, UAR) on MAFW and DFEW datasets. Bold indicates the best score. \* Measured using the official checkpoint provided by the authors.

Read the reasoning content and respond with the appropriate emotion in {Emotion List}.

Reply with only the emotion word.

Reason: {Explanation} Answer Emotion:

Figure 3: Prompt used to evaluate the emotion conveyed in the generated explanation. {Emotion List} is a permutation of the GT emotion label according to the evaluation dataset. {Explanation} is  $E_i$  from output  $o_i$ .

learning rate, a cosine scheduler, a 0.03 warmup ratio, and a batch size of 32. The ERV module is trained separately for 10 epochs with a batch size of 64 and a  $3e^{-5}$  learning rate. Those are trained on 8 NVIDIA RTX 3090 GPUs. For GRPO training, we use a batch size of 16 with 2 gradient accumulation steps, generating  $G = 16$  samples per input, and apply a KL divergence penalty of 0.04. For inference, we use a temperature of 0.3. This stage is conducted on 8 NVIDIA A100 GPUs.

## 5 Experimental Results

### 5.1 Main Results

To verify the effectiveness of our proposed Explanation reward in the quality of generated explanations, we evaluated on three proposed metrics: EEA, EPC, and FCR, with standard recognition accuracy WAR and UAR. Our model is compared with three models trained in different ways using the same multimodal emotion description dataset and HumanOmni architecture. (1) **EMER-SFT**: A supervised fine-tuned model trained on the EMER dataset for emotion reasoning, available in both 0.5B and 7B. (2) **R1-Omni**: A model trained with the GRPO using answer and format rewards. (3) **Baseline**: Since R1-Omni lacks training details and relies on an in-house dataset, we re-trained EMER-SFT with GRPO on the answer and format rewards using the same training dataset and hyperparameter settings as our method for a fair comparison.

As shown in Table 1, both EMER-SFT and R1-Omni exhibit low FCR, achieving only 29.67% and 31.80%, respectively. This indicates that although R1-Omni significantly improves emotion recognition accuracy on WAR, from 38.95%

to 57.68% compared to EMER-SFT, 44.87% of its predictions (almost half) are not supported by emotionally aligned explanations. Furthermore, its ability to independently produce explanations aligned with the target emotions (EEA) improves only marginally from 33.06% to 37.87%. In contrast, our proposed model trained with an Explanation reward not only achieves comparable recognition performance to R1-Omni but also demonstrates substantial improvements in both FCR and EEA. FCR increases from 29.67% to 43.57% and also EEA increases from 33.06% to 48.93%. A similar tendency is observed on the DFEW, where FCR improves from 32.63% to 53.75% and EEA improves from 35.14% to 59.45%.

We also compared our model with EMER-SFT (7B) and Emotion-LLaMA (7B), both of which use 7B-sized backbones but have different architectures. In the evaluation on the DFEW, both models exhibited low performance, achieving EEA scores of 44.86% and 41.15%, and FCR scores of 42.21% and 32.56%, respectively. These results indicate that SFT on existing multimodal video–reasoning pair datasets yields unreliable emotion reasoning performance regardless of the model architecture. In contrast, despite using a smaller 0.5B LLM size, our model outperforms both baselines in terms of FCR and EEA, while also achieving noticeably better emotion recognition accuracy.

About EPC evaluation, EMER-SFT shows significant performance compared to RL-based trained models. Namely, the final predicted answer is derived based on the generated explanations. However, when trained with GRPO on answer and format reward, this consistency deteriorates, as the model tends to focus solely on producing the final answer, leading to a degradation in EPC. In contrast, our proposed Explanation Reward enables the model to maintain this consistency between the explanation and the predicted answer, achieving performance that is comparable to or even surpasses that of EMER-SFT (70.66% vs. 68.58% in MAFW and 64.30% vs. 73.06% in DFEW). This indicates that our Explanation Reward effectively guides the model to generate explanations and predictions in a coherent and aligned manner.

### 5.2 Ablation Study

**Effectiveness of Emotional Rationale Verifier** To verify our proposed training strategy on the ERV module, we removed the data augmentation stage. (1) Baseline, without

Setting	EEA	FCR	EPC	WAR	UAR
Baseline	32.41	27.27	40.98	56.83	38.66
<b>ERV</b>					
W/o Data Aug.	46.88	42.62	65.46	56.78	38.60
<b>Explanation Reward (<math>R_E</math>)</b>					
W/o Neutral Filt.	46.28	42.13	64.70	57.16	38.67
W/o Sent. Verif.	47.54	43.17	68.36	56.78	38.66
Ours	<b>48.93</b>	<b>43.57</b>	<b>68.58</b>	<b>57.49</b>	<b>39.12</b>

Table 2: Ablation study of the ERV and  $R_E$  on MAFW. All values are reported in percentage (%).

Evaluation	Ours	R1-Omni
Win Rate	53.6%	46.4%
Assessment Score	3.52	3.42

Table 3: Human evaluation comparison between our model and R1-Omni

the ERV module, can’t reward on the generated explanation. It only focuses on the format and answer. (2) Without a data augmentation strategy, training the ERV solely on the EMER and MERR-fine datasets boosts EEA from 32.41% to 46.88% and FCR from 27.27% to 42.62%, while keeping WAR and UAR unchanged. (3) Adding the augmented data to ERV further raises EEA + 1.26 pp, FCR + 1.31 pp, EPC + 2.90 pp, and overall recognition accuracy. These results suggest that the ERV’s judgment capability is constrained by the size of its training data. Therefore, even when additional multimodal-emotion descriptions are unavailable, text-level data augmentation can noticeably enhance the model’s multimodal reasoning ability.

**Effectiveness of Explanation Reward** To validate the effectiveness of our novel proposed Explanation Reward, we compare it against a variant trained without our two key components: (1) Neutral Filtering and (2) Sentence-Level Verification. In this variant, the reward calculation does not filter out non-emotional sentences (as described in Equation 4 for the neutral case), and the ERV evaluates the emotional coherence of the entire explanation much more holistically. We observed a performance drop when each of these components was removed individually. Specifically neutral filtering, which excludes emotionally irrelevant sentences, was particularly effective in guiding the model to generate emotion-focused explanations.

### 5.3 Human Evaluation about Explanation Quality

Through the main results and ablation studies, we confirmed that explanations from our model better align with both target and predicted emotions. However, their quality in terms of how effectively they explain the emotions in the videos had not been directly evaluated. To address this, we conducted a human study comparing our model’s explanations to those from R1-Omni. Specifically, we randomly selected

Closed-source LLM	Model	EEA	FCR	EPC
GPT-4.1-mini	Baseline	32.41	27.27	40.98
	Ours	48.93	43.57	68.58
Gemini-Flash-2.5	Baseline	32.24	27.76	41.80
	Ours	48.52	42.73	65.96

Table 4: Performance of Baseline and Ours (EEA, FCR, EPC) using different closed-source LLMs. All values are reported in percentage (%).

14 samples from each model (total 28), where the generated emotion matched the ground-truth. Twenty human raters evaluated each explanation on a 1–5 scale based on how well it explained the corresponding video’s emotion. For each sample, we recorded which model received the higher score and computed the average score across all samples. As shown in Table 3, our model achieved a 53.6% win rate and a higher average score (3.52 vs. 3.42), indicating that it produces more informative and emotionally grounded explanations than R1-Omni.

### 5.4 Robustness of the Proposed Reasoning Evaluation Metrics

Since the evaluation of EEA, FCR, and EPC was conducted using a single closed-source LLM, we further validated the robustness of the proposed metrics by repeating the evaluation with a different model family, *Gemini-2.5-Flash* (Comanici et al. 2025). As shown in Table 4, the evaluation on MAFW reveals a consistent trend, with both the Baseline and our model exhibiting improved reasoning quality. This demonstrates that our metrics are independent of the specific LLM evaluator and consistently capture the alignment between reasoning and emotion across model families. Moreover, a human survey confirmed a 92.9% agreement between participants’ judgments on the conveyed emotion and those of GPT-4.1-mini.

## 6 Conclusion

We proposed training an MLLM for the MER task utilizing our proposed ERV and Explanation Reward, aiming to enhance the coherence between emotion recognition and explanation. Compared to prior approaches, where coherence between predictions and explanations among correctly predicted samples remained around half (55.13% on MAFW and 42.92% on DFEW), our method significantly improves this alignment, achieving 75.79% and 59.77%. Furthermore, human evaluations consistently confirmed that our approach produces more emotionally and contextually grounded emotion explanations for emotional video data. As a result, our work enables emotionally coherent and trustworthy explanations alongside accurate emotion recognition, representing a significant advancement toward emotionally intelligent human–computer interaction.

## Acknowledgments

This work was partly supported by two funds: Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2025-25442384, QuestioningAI: Development of Active Reasoning Cooperative Multimodal Agent Technology), and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2022R1A2C2005529). Additionally, this work was supported by the National Supercomputing Center with supercomputing resources including technical support (KSC-2025-CRE-0090).

## References

- Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J. N.; Lee, S.; and Narayanan, S. S. 2008. IEMOCAP: Interactive Emotional Dyadic Motion Capture Database. *Lang. Resour. Eval.*, 42: 335–359.
- Chen, L.; Li, L.; Zhao, H.; Song, Y.; and Vinci. 2025a. R1-V: Reinforcing Super Generalization Ability in Vision-Language Models with Less Than \$3. <https://github.com/Deep-Agent/R1-V>. Accessed: 2025-02-02.
- Chen, Y.; Ge, Y.; Wang, R.; Ge, Y.; Cheng, J.; Shan, Y.; and Liu, X. 2025b. GRPO-CARE: Consistency-Aware Reinforcement Learning for Multimodal Reasoning. arXiv:2506.16141.
- Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proc. CVPR*.
- Cheng, Z.; Cheng, Z.-Q.; He, J.-Y.; Wang, K.; Lin, Y.; Lian, Z.; Peng, X.; and Hauptmann, A. 2024a. Emotion-LLaMA: Multimodal Emotion Recognition and Reasoning with Instruction Tuning. In *NeurIPS*.
- Cheng, Z.; Leng, S.; Zhang, H.; Xin, Y.; Li, X.; Chen, G.; Zhu, Y.; Zhang, W.; Luo, Z.; Zhao, D.; et al. 2024b. VideoLLaMA 2: Advancing Spatial-Temporal Modeling and Audio Understanding in Video-LLMs. arXiv:2406.07476.
- Chu, Y.; Xu, J.; Yang, Q.; Wei, H.; Wei, X.; Guo, Z.; Leng, Y.; Lv, Y.; He, J.; Lin, J.; et al. 2024. Qwen2-Audio Technical Report. arXiv:2407.10759.
- Chumachenko, K.; Iosifidis, A.; and Gabbouj, M. 2024. MMA-DFER: MultiModal Adaptation of unimodal models for Dynamic Facial Expression Recognition in-the-wild. In *Proc. CVPR*.
- Comanici, G.; Bieber, E.; Schaekermann, M.; Pasupat, I.; Sachdeva, N.; Dhillon, I.; Blistein, M.; Ram, O.; and et al. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. arXiv:2507.06261.
- Demszky, D.; Movshovitz-Attias, D.; Ko, J.; Cowen, A.; Nemade, G.; and Ravi, S. 2020. GoEmotions: A Dataset of Fine-Grained Emotions. In *Proc. ACL*.
- Fei, H.; Zhang, H.; Wang, B.; Liao, L.; Liu, Q.; and Cambria, E. 2024. EmpathyEar: An Open-source Avatar Multimodal Empathetic Chatbot. In *Proc. ACL*.
- Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The Llama 3 Herd of Models. arXiv:2407.21783.
- Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv:2501.12948.
- Hu, Z.; Yuan, K.; Liu, X.; Yu, Z.; Zong, Y.; Shi, J.; Yue, H.; and Yang, J. 2025. Feallm: Advancing facial emotion analysis in multimodal large language models with emotional synergy and reasoning. In *Proc. ACM MM*.
- Jiang, X.; Zong, Y.; Zheng, W.; Tang, C.; Xia, W.; Lu, C.; and Liu, J. 2020. Dflew: A large-scale database for recognizing dynamic facial expressions in the wild. In *Proc. ACM MM*.
- Kim, Y.; Chung, J.; Kim, J.; Lee, S.; Lee, S.; Kim, J.; Yang, C.; and Yu, Y. 2025. Speaking Beyond Language: A Large-Scale Multimodal Dataset for Learning Nonverbal Cues from Video-Grounded Dialogues. arXiv:2506.00958.
- Kossaifi, J.; Tzimiropoulos, G.; Todorovic, S.; and Pantic, M. 2017. AFEW-VA database for valence and arousal estimation in-the-wild. *Image Vis. Comput.*
- Lee, H.; Kim, J.; Kim, H.; and Ro, Y. M. 2025. ReFoCUS: Reinforcement-Guided Frame Optimization for Contextual Understanding. arXiv:2506.01274.
- Lee, J.; Kim, S.; Kim, S.; Park, J.; and Sohn, K. 2019. Context-aware emotion recognition networks. In *Proc. ICCV*.
- Lee, S.; Mac Kim, S.; Kim, M.; Kang, D.; Yang, D.; Kim, H.; Kang, M.; Jung, D.; Kim, M.; Lee, S.; et al. 2024. Cactus: Towards Psychological Counseling Conversations using Cognitive Behavioral Theory. In *Proc. EMNLP*.
- Li, F.; Zhang, R.; Zhang, H.; Zhang, Y.; Li, B.; Li, W.; Ma, Z.; and Li, C. 2024. LLaVA-Next-Interleave: Tackling Multi-Image, Video, and 3D in Large Multimodal Models. arXiv:2407.07895.
- Li, X.; Yan, Z.; Meng, D.; Dong, L.; Zeng, X.; He, Y.; Wang, Y.; Qiao, Y.; Wang, Y.; and Wang, L. 2025. VideoChat-R1: Enhancing Spatio-Temporal Perception via Reinforcement Fine-Tuning. arXiv:2504.06958.
- Lian, Z.; Chen, H.; Chen, L.; Sun, H.; Sun, L.; Ren, Y.; Cheng, Z.; Liu, B.; Liu, R.; Peng, X.; et al. 2025. AffectGPT: A New Dataset, Model, and Benchmark for Emotion Understanding with Multimodal Large Language Models. In *ICML*.
- Lian, Z.; Sun, H.; Sun, L.; Chen, L.; Chen, H.; Gu, H.; Wen, Z.; Chen, S.; Zhang, S.; Yao, H.; et al. 2024. Open-Vocabulary Multimodal Emotion Recognition: Dataset, Metric, and Benchmark. arXiv:2410.01495.
- Lian, Z.; Sun, L.; Xu, M.; Sun, H.; Xu, K.; Wen, Z.; Chen, S.; Liu, B.; and Tao, J. 2023. Explainable Multimodal Emotion Reasoning. arXiv:2306.15401.
- Liu, Y.; Dai, W.; Feng, C.; Wang, W.; Yin, G.; Zeng, J.; and Shan, S. 2022. Mafw: A large-scale, multi-modal, compound affective database for dynamic facial expression recognition in the wild. In *Proc. ACM MM*.
- Liu, Z.; Sun, Z.; Zang, Y.; Dong, X.; Cao, Y.; Duan, H.; Lin, D.; and Wang, J. 2025. Visual-rft: Visual reinforcement fine-tuning. arXiv:2503.01785.
- Livingstone, S. R.; and Russo, F. A. 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLoS ONE*, 13(5): e0196391.
- Narayan, K.; VS, V.; Chellappa, R.; and Patel, V. M. 2025. Facexformer: A unified transformer for facial analysis. In *Proc. ICCV*.
- Park, S.; Kim, C.; Rha, H.; Kim, M.; Hong, J.; Yeo, J.; and Ro, Y. 2024. Let's Go Real Talk: Spoken Dialogue Model for Face-to-Face Conversation. In *Proc. ACL, 2024*.
- Park, S.; Kim, H.; Kim, J.; Kim, S.; and Ro, Y. M. 2025. DIP-R1: Deep Inspection and Perception with RL Looking Through and Understanding Complex Scenes. arXiv:2505.23179.

- Perepelkina, O.; Kazimirova, E.; and Konstantinova, M. 2018. RAMAS: Russian multimodal corpus of dyadic interaction for affective computing. In *Proc. SPECOM*. Springer.
- Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; and Mihalcea, R. 2019. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proc. ACL*.
- Schmidgall, S.; Ziaei, R.; Harris, C.; Reis, E.; Jopling, J.; and Moor, M. 2024. AgentClinic: A Multimodal Agent Benchmark to Evaluate AI in Simulated Clinical Environments. arXiv:2405.07960.
- Schutz, P. A.; and Pekrun, R. E. 2007. *Emotion in education*. Elsevier Academic Press.
- Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. arXiv:2402.03300.
- Shen, H.; Liu, P.; Li, J.; Fang, C.; Ma, Y.; Liao, J.; Shen, Q.; Zhang, Z.; Zhao, K.; Zhang, Q.; et al. 2025. Vlm-r1: A stable and generalizable r1-style large vision-language model. arXiv:2504.07615.
- Sun, L.; Lian, Z.; Liu, B.; and Tao, J. 2023. Mae-dfer: Efficient masked autoencoder for self-supervised dynamic facial expression recognition. In *Proc. ACM MM*.
- Sun, L.; Lian, Z.; Liu, B.; and Tao, J. 2024. Hicmae: Hierarchical contrastive masked autoencoder for self-supervised audio-visual emotion recognition. *Inf. Fusion*, 108: 102382.
- Team, Q. 2024. Qwen2 Technical Report. arXiv:2407.10671.
- Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288.
- Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. arXiv:2409.12191.
- Wang, Y.; Sun, Y.; Huang, Y.; Liu, Z.; Gao, S.; Zhang, W.; Ge, W.; and Zhang, W. 2022. Ferv39k: A large-scale multi-scene dataset for facial expression recognition in videos. In *Proc. CVPR*.
- Wei, T.; Yang, Y.; Xing, J.; Shi, Y.; Lu, Z.; and Ye, D. 2025. GTR: Guided Thought Reinforcement Prevents Thought Collapse in RL-Based VLM Agent Training. arXiv:2503.08525.
- Xing, B.; Yu, Z.; Liu, X.; Yuan, K.; Ye, Q.; Xie, W.; Yue, H.; Yang, J.; and Kälviäinen, H. 2024. Emo-llama: Enhancing facial emotion understanding with instruction tuning. arXiv:2408.11424.
- Xing, Z.; Hu, X.; Fu, C.-W.; Wang, W.; Dai, J.; and Heng, P.-A. 2025. EchoInk-R1: Exploring Audio-Visual Reasoning in Multimodal LLMs via Reinforcement Learning. arXiv:2505.04623.
- Yang, Q.; Bai, D.; Peng, Y.-X.; and Wei, X. 2025a. Omni-Emotion: Extending Video MLLM with Detailed Face and Audio Modeling for Multimodal Emotion Analysis. arXiv:2501.09502.
- Yang, Q.; Yao, S.; Chen, W.; Fu, S.; Bai, D.; Zhao, J.; Sun, B.; Yin, B.; Wei, X.; and Zhou, J. 2025b. HumanOmniV2: From Understanding to Omni-Modal Reasoning with Context. arXiv:2506.21277.
- Zadeh, A. B.; Liang, P. P.; Poria, S.; Cambria, E.; and Morency, L.-P. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proc. ACL*.
- Zhang, H.; Meng, Z.; Luo, M.; Han, H.; Liao, L.; Cambria, E.; and Fei, H. 2025. Towards multimodal empathetic response generation: A rich text-speech-vision avatar-based benchmark. In *Proc. WWW*.
- Zhang, S.; Sun, P.; Chen, S.; Xiao, M.; Shao, W.; Zhang, W.; Liu, Y.; Chen, K.; and Luo, P. 2024a. Gpt4roi: Instruction tuning large language model on region-of-interest. In *Proc. ECCV*.
- Zhang, S.; Tao, X.; Chuang, Y.; and Zhao, X. 2021. Learning deep multimodal affective features for spontaneous speech emotion recognition. *Speech Commun.*, 127: 73–81.
- Zhang, S.; Yang, Y.; Chen, C.; Zhang, X.; Leng, Q.; and Zhao, X. 2024b. Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects. *Expert Syst. Appl.*, 237: 121692.
- Zhang, S.; Zhang, S.; Huang, T.; and Gao, W. 2017. Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching. *IEEE Trans. Multimedia*, 20(6): 1576–1590.
- Zhao, F.; Li, M.; Xu, L.; Jiang, W.; Gao, J.; and Yan, D. 2025a. FaVChat: Unlocking Fine-Grained Facial Video Understanding with Multimodal Large Language Models. arXiv:2503.09158.
- Zhao, J.; Wei, X.; and Bo, L. 2025. R1-Omni: Explainable Omni-Multimodal Emotion Recognition with Reinforcement Learning. arXiv:2503.05379.
- Zhao, J.; Yang, Q.; Peng, Y.; Bai, D.; Yao, S.; Sun, B.; Chen, X.; Fu, S.; Wei, X.; Bo, L.; et al. 2025b. Humanomni: A large vision-speech language model for human-centric video understanding. arXiv:2501.15111.
- Zhao, Z.; and Liu, Q. 2021. Former-dfer: Dynamic facial expression recognition transformer. In *Proc. ACM MM*.