

# GRPO-Guided Modality Selection Enhanced LoRA-Tuned LLMs for Multimodal Emotion Recognition

Yang Chen<sup>1,2</sup>, Shuwan Yang<sup>1,2</sup>, Yan Xiang<sup>1,2\*</sup>, Ran Song<sup>1,2</sup>,  
Yuxin Huang<sup>1,2</sup>, Zhengtao Yu<sup>1,2</sup>

<sup>1</sup>Faculty of Information Engineering and Automation,  
Kunming University of Science and Technology, Kunming, China  
<sup>2</sup>Yunnan Key Laboratory of Artificial Intelligence, Kunming, China  
{cy\_xxx0, syang12}@stu.kust.edu.cn, sharonxiang@126.com,  
{song\_ransr, huangyuxin2004}@163.com, ztyu@hotmail.com

## Abstract

Multimodal emotion recognition in conversation (MERC) aims to identify speakers' emotional states by utilizing text, audio, and visual modalities. Although recent large language model (LLM)-based methods have demonstrated strong performance, they typically adopt static fusion strategies that integrate all available modalities uniformly. This overlooks the fact that the necessity of multimodal cues can vary significantly across utterances. In this work, we propose an adaptive modality selection framework for MERC. The core of our approach is a modality selection module based on Group Relative Policy Optimization (GRPO), which enables a LoRA-tuned LLM to reason about the necessity of multimodal input via chain-of-thought (CoT) generation. This process does not require manually labeled modality selection data and is trained in a fully unsupervised manner. The selected modality configuration is then provided as input to a downstream emotion classifier, which is also implemented using a LoRA-tuned LLM and trained to predict emotional states. Experimental results on benchmark multimodal dialogue datasets show that our method consistently outperforms strong baselines, demonstrating the effectiveness of adaptive modality selection in improving recognition accuracy. Our code is available at <https://github.com/youflyaway/Modality-Selection-Enhanced-LoRA-Tuned-LLMs>.

## 1 Introduction

Multimodal Emotion Recognition in Conversations (MERC) aims to identify the speaker's emotion by combining text, audio, and visual signals. Existing studies (Majumder et al., 2019; Li et al., 2022; Hu et al., 2022c; Li et al., 2023a) have primarily focused on multimodal feature fusion and alignment, yielding better results than unimodal baselines. Recently, Large Language Models (LLMs)

have achieved significant results in a variety of natural language processing tasks due to their powerful context modeling and reasoning capabilities. Some researchers have applied LLMs to emotion recognition tasks. InstructERC (Lei et al., 2023a) formulates emotion recognition as a generative task, introducing retrieval-based templates and emotion alignment mechanisms to model speaker interactions. VoiceERC (Wu et al., 2025) transforms speech features into natural language descriptions in a way that enables multimodal emotion analysis without architectural changes. DialogueLLM (Zhang et al., 2023) implements MERC using an end-to-end supervised instruction fine-tuning method by converting conversation videos into textual descriptions that will be used as complementary knowledge to construct high-quality instructions. Similarly, TextMI (Hasan et al., 2023) and (Richet et al., 2024) converts audio and visual inputs into structured textual descriptions, effectively reducing model complexity while preserving task performance.

Although LLM-based methods show advantages in multimodal emotion recognition tasks, they mainly use a fixed multimodal usage strategy, ignoring that different utterances may require different combinations of modalities. When there are semantic breaks or cryptic expressions in the conversation text, other modal cues in the historical conversation can effectively complement the key contextual information; while when the text contains sufficient emotion cues (e.g., explicit emotion words, intention indicators, or contextual markers), the introduction of multimodal features may cause the model to focus excessively on redundant information. As illustrated in Figure 1a, Joey's utterance in the Friends dataset is emotionally ambiguous in text alone, but visual and acoustic cues from earlier turns help clarify the emotional state. Conversely, in Figure 1b, the teacher's statement already provides sufficient semantic signals, allowing accurate

\*Corresponding author

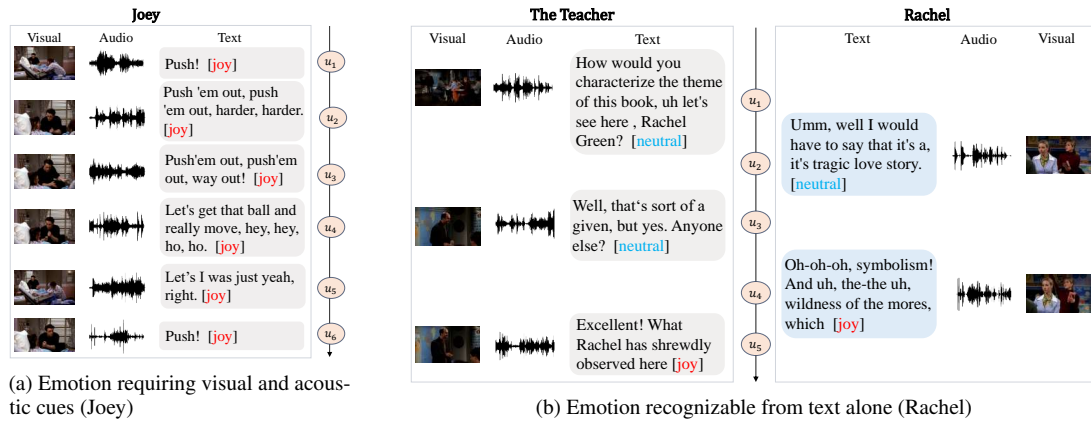


Figure 1: Examples of utterances with varying modality dependencies from the MELD dataset.

recognition based solely on the textual modality.

To address the limitations of fixed modality usage, we propose a MERC framework centered on adaptive modality selection. The goal is to dynamically determine whether to incorporate multimodal information based on contextual needs, thereby improving modality utilization. The method adopts a two-stage structure: in the first stage, the model learns in an unsupervised setting whether the current utterance requires multimodal cues from the historical conversation. We introduce a reinforcement learning mechanism based on Group Relative Policy Optimization (GRPO) (Shao et al., 2024) to train a LoRA-tuned (Hu et al., 2022b) LLM as a modality selector. Instead of using an explicit value function, the model leverages the average reward from multiple outputs for the same input as a learning signal. In the second stage, the selected unimodal or multimodal inputs are used to construct prompts for LoRA-based fine-tuning of the LLM on supervised emotion-labeled data, enabling the final emotion classification. The main contributions are as follows:

(1) We identify the limitation of fixed modality usage in MERC and propose a parameter-efficient two-stage framework that integrates modality selection with emotion recognition. By dynamically determining whether to incorporate multimodal information based on contextual needs, our method enables flexible adaptation to varying modality demands across different conversational scenarios, thereby enhancing emotion classification performance.

(2) We design a modality adaptive selection strategy based on GRPO, which guides LLM to learn when to introduce multimodal information under unlabeled conditions, and improves the model’s

ability to perceive the differences in modal demands in different contexts.

(3) Experiments on benchmark datasets demonstrate that our framework consistently outperforms competitive baselines across core metrics, validating the effectiveness and generalizability of the proposed adaptive modality selection mechanism.

## 2 Related Work

### 2.1 Emotion Recognition in Conversation

Following over a decade of development, numerous works have emerged in the domain of Emotion Recognition in Conversation (ERC). These works can be broadly categorized into four groups: Approaches that have been developed include transformers-based, GNN-based, recurrent-based and large language models methods.

Transformer-based approaches (Li et al., 2020; Song et al., 2022; Liu et al., 2023; Chudasama et al., 2022) seek to capture long-range emotional correlations in conversation by either adopting the vanilla Transformer architecture or modifying its self-attention blocks to conversational data. GNN-based works (Ghosal et al., 2019; Ishiwatari et al., 2020; Shen et al., 2021; Li et al., 2023b) use graphs to model human interactions in conversational scene as well as the effects between different modalities, i.e. speakers, utterances and multimodal features. Recurrent-based works (Hu et al., 2023; Lei et al., 2023b; Hazarika et al., 2018; Majumder et al., 2019) rely on LSTM or GRU-based encoders, often augmented with attention mechanisms, to track speaker-specific emotional state while maintaining a coherent view of the global conversation context.

## 2.2 Large Language Models for Emotion Recognition

LLMs have rapidly migrated from general-purpose NLP to specialized downstream tasks, including ERC. InstructERC (Lei et al., 2023a) prompts LLMs to predict emotion labels through instruction tuning, introducing a retrieval template module and emotional alignment tasks that grounds generation in conversational context. BiosERC (Xue et al., 2024) enriches the prompt with the speaker’s biographical information automatically distilled by LLMs, helping better understand emotional interactions in the conversation. CKERC (Fu, 2024) design prompts to generate interlocutors’ common-sense based on historical utterances with large language model. VoiceERC (Wu et al., 2025) incorporates emotional cues from speech into the model input for a more comprehensive understanding of the speaker’s emotional state.

## 3 Method

The proposed method follows a two-stage framework, as illustrated in Figure 2. First, the GRPO-guided modality selector determines whether multimodal cues from the historical context are necessary for the current utterance in an unsupervised manner. Second, a LoRA-tuned LLM performs emotion classification based on the selected input configuration (text-only or multimodal). Prior to these two stages, we transform raw acoustic and visual inputs into structured textual descriptions, ensuring compatibility with the LLM-based prompting paradigm.

### 3.1 Problem Formulation

Assume the multimodal dataset contains multiple conversations, each consisting of  $K$  utterances. For the  $k$ -th utterance, we define its associated multimodal input as:

$$M_k = \{u_k, \text{audio}_k, \text{video}_k, y_k\} \quad (1)$$

where  $u_k$  denotes the text of the  $k$ -th utterance,  $\text{audio}_k$  and  $\text{video}_k$  represent the corresponding raw audio and visual inputs respectively, and  $y_k$  is the associated emotion label.

The task consists of two sub-objectives:

(1) Modality Selection: Learn a policy  $\pi_\theta$  to decide whether to incorporate audio and visual description in addition to the text input for emotion recognition.

(2) Emotion Recognition: Train a classifier to predict the emotion label.

### 3.2 Multimodal Feature Processing

To unify all input modalities into a format suitable for LLM-based processing, we convert the raw audio and visual signals into structured textual descriptions using pretrained modality-specific encoders. These descriptions are then integrated with textual content to support downstream modeling.

For each utterance  $u_k$ , we process its associated audio segment  $\text{audio}_k$  using Qwen2-Audio (Chu et al., 2024), which generates a textual description aligned with the speech content:

$$C_{\text{aud}}^{(k)} = \text{Qwen\_Audio}(\text{prompt}, \text{audio}_k) \quad (2)$$

Here,  $C_{\text{aud}}^{(k)}$  captures emotional cues such as tone, pitch, and rhythm that complement the textual utterance.

Each utterance  $u_k$  is also assigned a visual input  $\text{video}_k$ . We employ Qwen2-VL (Wang et al., 2024) to generate a structured textual description of the video:

$$C_{\text{vid}}^{(k)} = \text{Qwen\_VL}(\text{prompt}, \text{video}_k) \quad (3)$$

This output captures high-level scene semantics and character interactions.

### 3.3 GRPO-Guided Modality Selection

To dynamically determine whether to incorporate multimodal information, we formulate modality selection as a binary reasoning task. Given the historical context of the current utterance, the model must decide whether additional audio or visual modalities are needed to enhance emotion understanding. We implement the modality decision model using an LLM with LoRA adapters (Hu et al., 2022b), which generates structured natural language responses indicating the necessity of multimodal information. The model is prompted with the current utterance, its dialogue history, and modality descriptions.

Since no ground-truth labels are available for modality usage, we employ GRPO (Shao et al., 2024)—a reinforcement learning method that optimizes the generation policy by sampling multiple candidate responses and assigning rewards based on carefully designed reward functions, thus avoiding the need for explicit supervision. GRPO allows the model to learn CoT reasoning and modality selection without requiring ground-truth. It explores

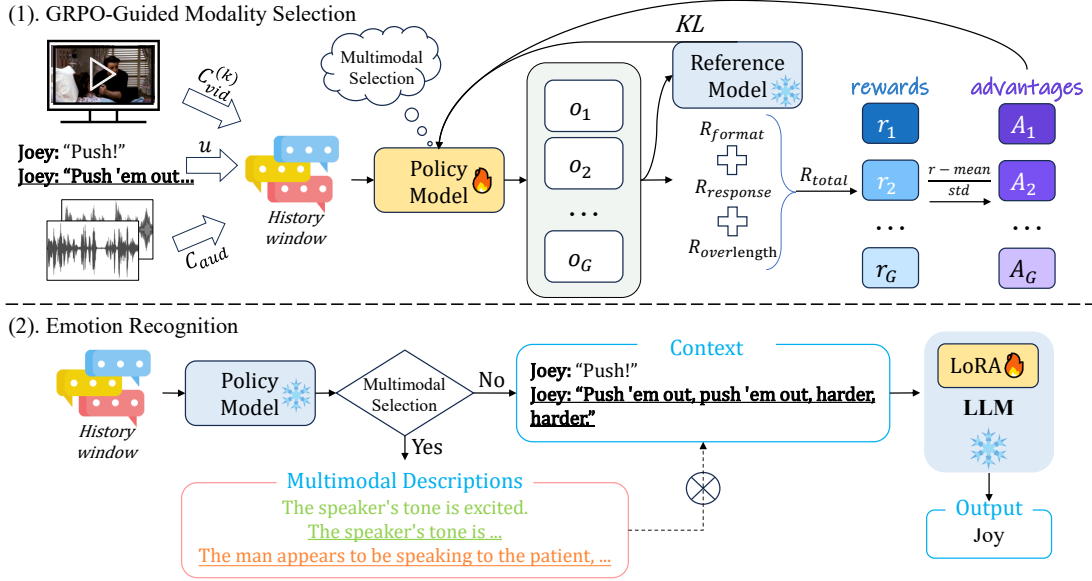


Figure 2: The proposed two-phase framework: (1) GRPO-Guided Modality Selection learns whether to use multimodal context; (2) LoRA-tuned LLM performs emotion recognition based on the selected modality configuration.

multiple reasoning paths, receives feedback via well-designed rewards, and gradually converges to better strategies. Compared to other RL algorithms like PPO (Schulman et al., 2017), GRPO does not require training an additional value model, which simplifies optimization and reduces overhead.

### 3.3.1 Question Definition

We frame the modality selection process as a prompt-based reasoning task, allowing LLMs to decide whether multimodal inputs are necessary. Specifically, the prompt  $q$  consists of three components: a natural language instruction that describes modality selection question, a decision flow, and a historical dialogue content that includes both textual and modality descriptions. It is defined as:

$$q = \{\text{Question, Decision Flow, Context}\},$$

$$\text{Context} = \{(u_1, C_{\text{aud}}^{(1)}), \dots, (u_k, C_{\text{aud}}^{(k)}), C_{\text{vid}}^{(k)}\} \quad (4)$$

Here,  $q$  is a fixed instruction designed to elicit reasoning about whether multimodal inputs are required for the target utterance (see Figure 8 in Appendix A.1).

By embedding textual, auditory, and video information in a unified template, the prompt enables model to jointly consider heterogeneous cues, thereby enhancing the reliability of modality selection.

### 3.3.2 Response Sampling and Group Rewards

The optimization process involves two models: a trainable policy model  $\pi_\theta$  and a frozen reference model  $\pi_{\text{ref}}$ , both initialized with the same parameters. During training,  $\pi_{\text{ref}}$  serves as a stable baseline to constrain the learning dynamics of  $\pi_\theta$ .

For each prompt  $q$ , we sample a group of responses  $o = \{o_1, o_2, \dots, o_G\}$  from the previous version of the policy  $\pi_{\text{old}}$ , where  $\pi_{\text{old}} = \pi_\theta$  before the latest update. These responses are grouped into  $G$  sets, each with potentially different lengths  $|o_i|$ .

Each response  $o_i$  is scored using a composite reward function, producing group-wise rewards  $r = \{r_1, r_2, \dots, r_G\}$ . The overall reward  $R(o_i)$  considers three components: format consistency, semantic relevance, and output length control, detailed as follows.

**Format Reward Function.** To encourage a consistent reasoning structure, we require each response to follow a predefined format: `<think>... </think><answer>... </answer>`. If  $o$  conforms to this format, we assign a reward of 1; otherwise, 0:

$$R_{\text{format}}(o) = \begin{cases} 1, & \text{if } o \text{ follows the format} \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

**Response Reward Function.** Since no ground-truth labels exist for modality requirements, we adopt a format-aligned reward function based on a predefined output template. Specifically, if the

model’s response strictly adheres to the target structure, a positive reward is issued; otherwise, it is penalized:

$$R_{\text{response}}(o) = \begin{cases} 1, & \text{if } o \text{ contains prompt} \\ & \text{aligned response to } q \\ -0.5, & \text{otherwise} \end{cases} \quad (6)$$

This reward is not based on detecting emotion- or modality-related semantic keywords. Instead, it is a format-constrained reward that checks whether the model’s response strictly adheres to a predefined template. Specifically, the model is prompted to follow this structure:

- *Is it recommended that the Caption be retained?(Yes, the captions should remain / No, the captions should be removed)*
- *Brief reasons why the Caption should be retained or not.*

A positive reward is issued only if the model’s response contains exactly one of the following phrases: "Yes, the captions should remain" or "No, the captions should be removed." Any other phrasing even if semantically correct is penalized. This pattern-based reward serves to enforce format regularity in outputs, especially in the early reinforcement learning phase. As training progresses, the model converges toward stable use of the expected pattern.

**Overlength Reward Function.** Using only response reward may encourage the model to stop reasoning early, once it reaches a seemingly valid but shallow answer (Liu et al., 2025; Yu et al., 2025). To mitigate this, we introduce an overlength penalty that discourages both excessively short and overly long outputs.

We define two length thresholds:  $L_{\text{cache}}$  (desired minimal reasoning length) and  $L_{\text{max}}$  (maximum acceptable length). Let  $|o|$  denote the token length of the generated response  $o$ . The overlength reward function is defined as:

$$R_{\text{overlength}}(o) = \begin{cases} 0, & |o| \leq L_{\text{cache}} \\ \frac{L_{\text{max}} - |o|}{L_{\text{max}} - L_{\text{cache}}}, & L_{\text{cache}} < |o| \leq L_{\text{max}} \\ -1, & |o| > L_{\text{max}} \end{cases} \quad (7)$$

This function penalizes overly long responses while allowing flexibility within a length tolerance interval.

The total reward for each response  $o$  is then computed as the unweighted sum of three components:

$$R_{\text{total}}(o) = R_{\text{format}}(o) + R_{\text{response}}(o) + R_{\text{overlength}}(o) \quad (8)$$

These reward functions jointly guide the model to generate well-structured, semantically accurate, and appropriately long responses when reasoning about whether multimodal features are necessary.

### CoT-inspired prompting Modality Reasoning.

To assess the necessity of multimodal information, we adopt a Chain-of-Thought (CoT)-inspired prompting strategy (Wei et al., 2022). Rather than relying on open-ended reasoning, our method embeds a step-by-step Decision Flow into the input prompt. As illustrated in Figure 9 in Appendix A.1.

By combining decision structure and natural language reasoning, our CoT-style prompting strategy not only enhances interpretability, but also improves the accuracy and robustness of adaptive modality selection. We empirically validate the effectiveness of this mechanism in the experiments that follow.

### 3.3.3 Advantage Estimation and Policy Updating

To train the modality selection policy, we adopt a group-based optimization strategy. For each prompt  $q$ , we sample a group of responses  $\{o_1, o_2, \dots, o_G\}$  from the current policy  $\pi_\theta$  (or its frozen copy  $\pi_{\theta_{\text{old}}}$ ). Each response  $o_i$  is assigned a scalar reward  $r_i$ , and the group-wise normalized advantage for token  $t$  in  $o_i$  is estimated as:

$$\hat{A}_{i,t} = \frac{r_i - \text{mean}(r)}{\text{std}(r)} \quad (9)$$

The policy is then updated using a clipped objective to stabilize learning:

$$\mathcal{J}^*(\theta) = \mathbb{E}_{o \sim \pi_\theta} \left[ \min(\text{ratio} \cdot \hat{A}, \text{clipped}) - \beta D_{\text{KL}}[\pi_\theta \| \pi_{\text{ref}}] \right] \quad (10)$$

Here, *ratio* denotes the likelihood ratio between the current and old policy at each token, and  $D_{\text{KL}}$  is the token-level KL divergence from the frozen reference model  $\pi_{\text{ref}}$  to prevent policy drift. The old policy  $\pi_{\text{old}}$  is periodically updated to match  $\pi_\theta$ . Here,  $\pi_\theta$  denotes the LoRA-tuned LLM, and the optimization is applied only to the LoRA modules.

This training strategy equips the model with chain-of-thought reasoning ability, allowing it to deliberate over each historical conversation window and decide whether multimodal cues are needed.

### 3.4 LoRA-Tuned LLMs for Emotion Recognition

In the second phase, we perform emotion classification using an LLM fine-tuned with LoRA for efficient parameter adaptation. Based on the output of the modality selection stage, the LLM receives either a text-only prompt or a prompt augmented with structured audio and visual descriptions. The input follows a unified template, as shown in Figure 10 in Appendix A.1, where modality-specific fields are conditionally included. Formally, the prompting input  $x$  is defined as:

$$x = \begin{cases} \text{prompting}(u), & \text{if multimodal} \\ \text{prompting}(u, C_{\text{aud}}, C_{\text{vid}}), & \text{selection is No} \\ & \text{otherwise} \end{cases} \quad (11)$$

This setup allows the model to flexibly adapt to different modality configurations while maintaining low training cost. We fine-tune the LLM using LoRA to predict the target emotion label  $\hat{y}_k$  given the input prompt  $x$ . The model is optimized with standard cross-entropy loss between predictions and ground-truth labels. At inference time, the model directly generates the emotion label based on the constructed prompt.

## 4 Experiments and Results

### 4.1 Dataset

We conduct experiments on two benchmarks: MELD (Poria et al., 2018) and IEMOCAP (Busso et al., 2008), covering 6–7 emotion categories. Dataset statistics are provided in the Appendix A.2 (Table 3).

### 4.2 Baselines

We compare our method against two categories of baselines: (1) including DialogueRNN (Majumder et al., 2019), MMGCN (Hu et al., 2021), DialogueTRM (Mao et al., 2021), MM-DFN (Hu et al., 2022a), EmoCaps (Li et al., 2022), and MPT-HCL (Zou et al., 2023), which are built on transformer-, recurrent-, or GNN-based encoders with multimodal fusion modules; and (2) including InstructERC (Lei et al., 2023a), BiosERC (Xue

et al., 2024), and VoiceERC (Wu et al., 2025), which leverage LLM-based backbones. Details and experiment settings are provided in Appendix A.3 and Appendix A.4.

### 4.3 Main Results

The experimental results are shown in Tables 1. We report both accuracy (ACC) and weighted F1 score (w-F1) to evaluate model performance. Our proposed method achieves the best overall performance, surpassing all fine-tuned LLM baselines on both datasets. Specifically, it improves w-F1 by 1.65% on IEMOCAP and 0.94% on MELD compared to InstructERC, demonstrating the effectiveness of adaptive multimodal integration guided by our two-stage framework. On the MELD dataset, fine-tuned LLMs consistently outperform traditional models based on BERT, Transformer, or GNN architectures. This performance gap may stem from MELD’s more diverse and open-domain conversational scenarios, where LLMs better leverage their reasoning capabilities. In contrast, on the laboratory-recorded IEMOCAP dataset with richer and more structured multimodal cues, pre-trained models perform competitively, highlighting the benefit of detailed audio-visual information in controlled settings.

We also observe that VoiceERC, which incorporates fixed multimodal descriptions into LLMs, performs slightly better than InstructERC on IEMOCAP but underperforms on MELD. One possible explanation is that directly introducing multimodal information may not always be beneficial—especially when textual content is already sufficient or when irrelevant modalities introduce noise. In contrast, our adaptive selection mechanism allows the model to leverage multimodal cues more selectively and effectively, leading to consistent improvements.

### 4.4 Ablation Study

We conduct ablation studies to assess the contribution of key components in our framework. Specifically, we consider the following variants:

- **w/o GRPO & modality desc:** Both the selection policy and multimodal inputs are removed. The model is thus reduced to standard text-only reasoning over the current window, without any modality adaptation.
- **w/o GRPO:** We disable the GRPO-based modality selection while retaining multimodal descriptions for each conversation window.

Model	IEMOCAP		MELD	
	ACC	w-F1	ACC	w-F1
DialogueRNN	69.38	69.37	66.70	65.31
MMGCN	69.62	69.61	66.40	65.21
DialogueTRM	69.87	69.91	66.70	65.76
MM-DFN	69.87	69.91	66.55	65.48
EmoCaps	–	71.77	–	64.00
MPT-HCL	72.83	72.51	65.86	65.02
TelME	–	70.48	–	67.37
InstructERC	–	71.39	–	69.15
BiosERC	–	69.02	–	68.72
VoiceERC	–	72.59	–	67.60
<b>Ours</b>	<b>72.95</b>	<b>73.04</b>	<b>71.00</b>	<b>70.09</b>

Table 1: ACC and w-F1 scores of different models on the IEMOCAP and MELD datasets.

This prevents the model from dynamically deciding whether to incorporate additional modality information.

- **w/o overlength reward:** The overlength penalty is excluded during modality policy training. This allows us to isolate the effect of reasoning length control on selection quality.

From Table 2, we can observe that removing GRPO leads to a notable performance drop, indicating that dynamic modality selection is critical for context-aware emotion recognition. When textual input already conveys rich affective content, multimodal signals may become redundant or even distracting. Conversely, in semantically ambiguous contexts, multimodal cues provide complementary information that enhances prediction accuracy.

Performance degrades further when both GRPO and modality descriptions are removed, suggesting that structured multimodal cues are beneficial. However, this benefit is highly context-dependent. As shown by the performance of VoiceERC, which incorporates full multimodal information without selection, addition of modality inputs may not always help. This reinforces the importance of adaptive and context-aware modality integration.

The overlength reward mainly constrains the response length to encourage balanced reasoning traces. Its removal leads to a moderate drop in performance, as the policy still leverages GRPO and contextual inputs, though with increased variance in output quality.

Method	IEMOCAP	MELD
Ours	<b>73.04</b>	<b>70.09</b>
w/o GRPO	72.00	69.57
w/o GRPO & modality desc.	71.32	69.23
w/o overlength reward	72.26	69.65

Table 2: W-F1 scores under different ablation settings on IEMOCAP and MELD.

#### 4.5 Impact of Historical Window Length

We analyze the impact of historical window length on model performance. As shown in Figure 3, the blue curve represents our method, which selectively incorporates multimodal information based on contextual relevance, while the red curve corresponds to a uniform strategy that indiscriminately adds multimodal inputs to all windows. Our adaptive strategy consistently outperforms the fixed integration baseline across all window lengths. The performance gap becomes more pronounced as the window length increases, indicating that selective integration is particularly beneficial in longer contexts, where irrelevant or redundant cues may otherwise impair emotion understanding. Moreover, we observe that medium-length windows (10–15 utterances) yield the best overall performance, suggesting that a balanced context span is beneficial.

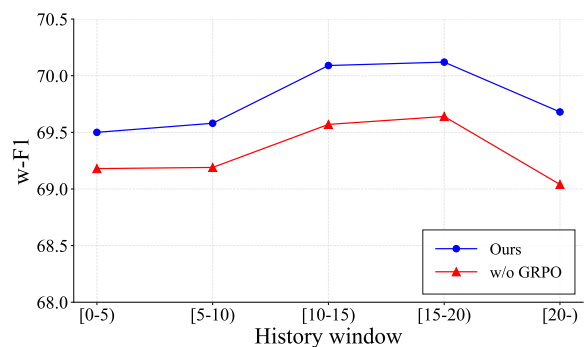


Figure 3: Performance comparison across different history window lengths on MELD.

#### 4.6 Fine-Grained Emotion Analysis

Figure 4 depicts a fine-grained comparison of four methods: EmoCaps, MPT-HCL, TelME, and Ours across the seven emotion categories on MELD, using per-class F1 as the evaluation metric. The black polygon representing our method consistently encloses the largest area, indicating both superior average performance and a more balanced distribution across emotion types.

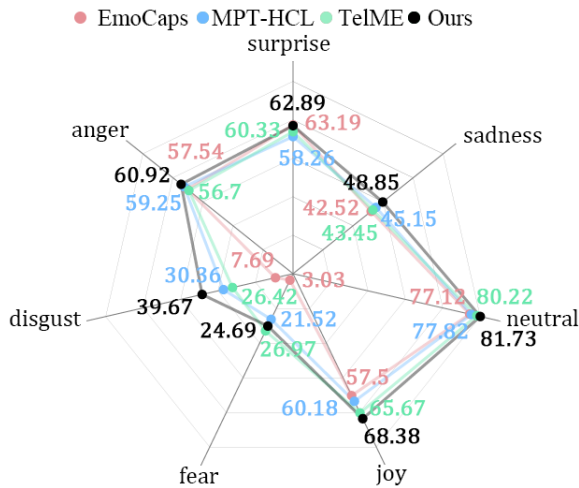


Figure 4: Performance of different methods on MELD

#### 4.7 Impact of Backbone Language Model

To isolate the effect of the backbone language model, we substitute only the LLM while keeping all other modules unchanged, and report the resulting w-F1 scores on MELD (Figure 5):

Incorporating full multimodal descriptions into every conversation window yields consistent absolute improvements of 0.5–0.9% w-F1 over the text-only baseline. Building on this, our GRPO-based selective integration strategy further improves performance by 0.4–0.8% W-F1, demonstrating that targeted multimodal integration effectively balances informativeness and redundancy. Among all backbones, LLaMA3-8B (Grattafiori et al., 2024) attains the highest absolute scores, followed by GLM4-9B (GLM et al., 2024) and then Qwen2-7B (Yang et al., 2024a), each exhibiting comparable relative gains.

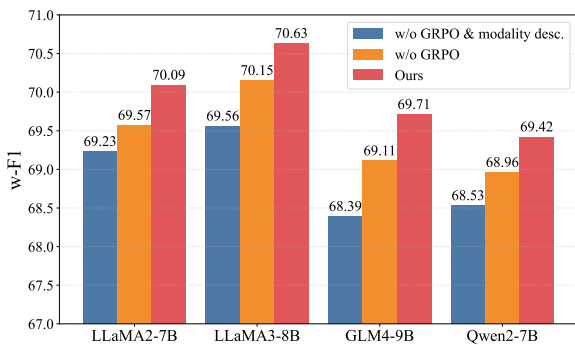


Figure 5: Performance comparison of different base model on MELD

#### 4.8 Comparison of text-only and vision-language backbone models

To evaluate the impact of introducing vision-language backbone models, we compare four large models under identical fine-tuning settings on the MELD dataset: LLaMA2-7B (Touvron et al., 2023), Qwen2-7B (Yang et al., 2024a), Qwen2-VL-7B (Wang et al., 2024), and MiniCPM-V-2.6-8B (Yao et al., 2024), and report their w-F1 scores (Figure 6). All models are fed with the same conversation transcripts; additionally, the vision-language models (Qwen2-VL-7B, MiniCPM-V-2.6-8B) receive aligned video clips corresponding to the target utterance. Owing to GPU constraints, we provide only the clip corresponding to the target utterance.

We observe that the text-only LLMs consistently outperform their vision-language counterparts. This suggests that at the 7B scale, introducing raw visual inputs through vision-language models may dilute the model’s capacity for effective conversation-level reasoning. A likely explanation is that processing video requires allocating model parameters to vision encoding, thereby reducing resources available for language understanding, which is critical for this task. These results highlight that simply switching to vision-language backbones does not guarantee better performance. Instead, they reinforce the need for selective and context-aware multimodal integration, rather than unconditional fusion.

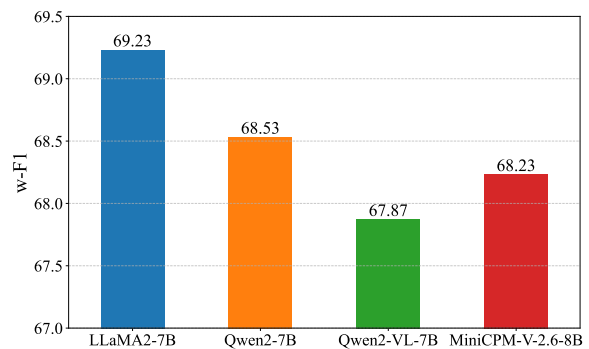


Figure 6: Comparison of text-only and vision-language backbone models on MELD

#### 4.9 Error Analysis

Although our method achieves excellent performance, it still fails to detect certain emotion categories. We analyzed the confusion matrices of the test set on MELD. As shown in Figure 7, it can be seen that (1) Joy is often misclassified as surprise



or neutral, and vice versa. Similarly, disgust and anger are frequently confused. This suggests that emotions with overlapping arousal characteristics are harder to distinguish, even for LLMs with multimodal cues. (2) MELD contains an imbalanced distribution of emotion categories. Less frequent categories such as fear and disgust show lower prediction accuracy due to limited training signals and greater overlap with neighboring emotional states.

	surprise	sadness	neutral	joy	fear	disgust	anger
anger	27	15	69	15	5	16	198
disgust	4	1	25	0	0	24	14
fear	8	4	18	3	10	2	5
joy	24	7	90	266	3	1	11
neutral	43	24	1087	58	9	6	29
sadness	12	85	73	7	2	3	26
surprise	183	4	42	27	2	1	22

Figure 7: The confusion matrices of the test set on MELD. The rows and columns represent predicted and true labels respectively.

#### 4.10 Case Study

We present two cases (visualized in Appendix A.4, Figure 11) to illustrate the robustness of our method. In the first example, the utterance “Push!” contains limited semantic information in text. Our method successfully integrates audio-visual cues to infer the speaker’s joyful encouragement. In contrast, when textual cues already clearly convey affect (e.g., “Excellent!” after a correct answer), multimodal addition causes over-reliance on visually salient but redundant cues, leading to misclassification. Our approach selectively filters unnecessary modalities, yielding accurate predictions in both cases.

These cases confirm that GRPO-guided selection improves robustness by introducing multimodal features only when beneficial, while avoiding the potential noise introduced by indiscriminate fusion.

## 5 Conclusion

In this paper, we propose a novel two-stage framework for MERC, which integrates structured multimodal descriptions and adaptively determines their

necessity for each input. Unlike prior works that apply fixed fusion strategies, our method selectively incorporates audio and visual cues based on contextual relevance, improving the robustness and interpretability of emotion recognition. Experiments on two benchmark datasets demonstrate that our approach achieves competitive performance. Moreover, the framework is simple and readily applicable to a wide range of dialogue understanding tasks.

## Limitations

Despite the strong performance of our method, it has two limitations. First, due to computational constraints, we evaluate only models with fewer than 10B parameters. Second, our current method adopts a pipeline framework, which decouples modality selection and emotion classification. While this design offers modular interpretability and flexible adaptation, it may limit end-to-end optimization, which is worth exploring in future work.

## Acknowledgments

This work is supported by Key basic research project of Yunnan Province (Grant NO.202501AS070147), National Natural Science Foundation of China (Grant NO.62162037, 62266027, U21B2027, 62266028), Yunnan provincial major science and technology special plan projects (Grant NO.202302AD080003), Yunnan Province Young and Middle-aged Academic and Technical Leaders Reserve Talent Program (Grant NO: 202305AC160063)

## References

- Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeanette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359.
- Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, and 1 others. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.
- Vishal Chudasama, Purbayan Kar, Ashish Gudmalwar, Nirmesh Shah, Pankaj Wasnik, and Naoyuki Onoe. 2022. M2fnet: Multi-modal fusion network for emotion recognition in conversation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4652–4661.

- Yumeng Fu. 2024. Ckerc: Joint large language models with commonsense knowledge for emotion recognition in conversation. *arXiv preprint arXiv:2403.07260*.
- Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. Dialoguecn: A graph convolutional neural network for emotion recognition in conversation. *arXiv preprint arXiv:1908.11540*.
- Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, and 1 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Md Kamrul Hasan, Md Saiful Islam, Sangwu Lee, Wasifur Rahman, Iftekhar Naim, Mohammed Ibrahim Khan, and Ehsan Hoque. 2023. Textmi: Textualize multimodal information for integrating non-verbal cues in pre-trained language models. *arXiv preprint arXiv:2303.15430*.
- Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. 2018. Icon: Interactive conversational memory network for multimodal emotion detection. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2594–2604.
- Dou Hu, Yanan Bao, Lingwei Wei, Wei Zhou, and Songlin Hu. 2023. Supervised adversarial contrastive learning for emotion recognition in conversations. *arXiv preprint arXiv:2306.01505*.
- Dou Hu, Xiaolong Hou, Lingwei Wei, Lianxin Jiang, and Yang Mo. 2022a. Mm-dfn: Multimodal dynamic fusion network for emotion recognition in conversations. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7037–7041. IEEE.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022b. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.
- Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. 2022c. Unimse: Towards unified multimodal sentiment analysis and emotion recognition. *arXiv preprint arXiv:2211.11256*.
- Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin Jin. 2021. Mmgcn: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. *arXiv preprint arXiv:2107.06779*.
- Taichi Ishiwatari, Yuki Yasuda, Taro Miyazaki, and Jun Goto. 2020. Relation-aware graph attention networks with relational position encodings for emotion recognition in conversations. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 7360–7370.
- Shanglin Lei, Guanting Dong, Xiaoping Wang, Keheng Wang, and Sirui Wang. 2023a. Instructerc: Reforming emotion recognition in conversation with a retrieval multi-task llms framework. *CoRR*.
- Shanglin Lei, Xiaoping Wang, Guanting Dong, Jiang Li, and Yingjian Liu. 2023b. Watch the speakers: A hybrid continuous attribution network for emotion recognition in conversation with emotion disentanglement. In *2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 881–888. IEEE.
- Bobo Li, Hao Fei, Lizi Liao, Yu Zhao, Chong Teng, Tat-Seng Chua, Donghong Ji, and Fei Li. 2023a. Revisiting disentanglement and fusion on modality and context in conversational multimodal emotion recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5923–5934.
- Jiang Li, Xiaoping Wang, Guoqing Lv, and Zhi-gang Zeng. 2023b. Graphcfc: A directed graph based cross-modal feature complementation approach for multimodal conversational emotion recognition. *IEEE Transactions on Multimedia*, 26:77–89.
- Jingye Li, Meishan Zhang, Donghong Ji, and Yijiang Liu. 2020. Multi-task learning with auxiliary speaker identification for conversational emotion recognition. *arXiv preprint arXiv:2003.01478*.
- Zaijing Li, Fengxiao Tang, Ming Zhao, and Yusen Zhu. 2022. Emocaps: Emotion capsule based model for conversational emotion recognition. *arXiv preprint arXiv:2203.13504*.
- Xiao Liu, Jian Zhang, Heng Zhang, Fuzhao Xue, and Yang You. 2023. Hierarchical dialogue understanding with special tokens and turn-level attention. *arXiv preprint arXiv:2305.00262*.
- Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. 2025. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*.
- Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6818–6825.
- Yuzhao Mao, Guang Liu, Xiaojie Wang, Weiguo Gao, and Xuan Li. 2021. DialogueTRM: Exploring multimodal emotional dynamics in a conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2694–2704, Punta Cana, Dominican Republic. Association for Computational Linguistics.

- Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.
- Nicolas Richet, Soufiane Belharbi, Haseeb Aslam, Meike Emilie Schadt, Manuela González-González, Gustavo Cortal, Alessandro Lameiras Koerich, Marco Pedersoli, Alain Finkel, Simon Bacon, and 1 others. 2024. Textualized and feature-based models for compound multimodal emotion recognition in the wild. *arXiv preprint arXiv:2407.12927*.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and 1 others. 2024. Deepseek-math: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.
- Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. 2021. Directed acyclic graph network for conversational emotion recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1551–1560, Online. Association for Computational Linguistics.
- Xiaohui Song, Longtao Huang, Hui Xue, and Songlin Hu. 2022. Supervised prototypical contrastive learning for emotion recognition in conversation. *arXiv preprint arXiv:2210.08713*.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Zehui Wu, Ziwei Gong, Lin Ai, Pengyuan Shi, Kaan Donbekci, and Julia Hirschberg. 2025. Beyond silent letters: Amplifying LLMs in emotion recognition with vocal nuances. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2202–2218, Albuquerque, New Mexico. Association for Computational Linguistics.
- Jieying Xue, Minh-Phuong Nguyen, Blake Matheny, and Le-Minh Nguyen. 2024. Bioserc: Integrating biography speakers supported by llms for erc tasks. In *International Conference on Artificial Neural Networks*, pages 277–292. Springer.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 39 others. 2024a. Qwen2 technical report. *ArXiv*, abs/2407.10671.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024b. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.
- Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, and 1 others. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.
- Qiyang Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, and 1 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.
- Taeyang Yun, Hyunkuk Lim, Jeonghwan Lee, and Min Song. 2024. TelME: Teacher-leading multimodal fusion network for emotion recognition in conversation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 82–95, Mexico City, Mexico. Association for Computational Linguistics.
- Yazhou Zhang, Mengyao Wang, Youxi Wu, Prayag Tiwari, Qiuchi Li, Benyou Wang, and Jing Qin. 2023. DialogueLLM: Context and emotion knowledge-tuned large language models for emotion recognition in conversations. *arXiv preprint arXiv:2310.11374*.
- Shihao Zou, Xianying Huang, and Xudong Shen. 2023. Multimodal prompt transformer with hybrid contrastive learning for emotion recognition in conversation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5994–6003.

## A Appendix

### A.1 Prompt

**Question Prompt** As illustrated in Figure 8, each conversation window is accompanied by a three-part multimodal necessity prompt constructed:

- **Question.** A fixed instruction that prompts the model to determine whether multimodal descriptions are necessary for classifying the emotion of the target utterance.
- **Decision Flow.** A step-by-step reasoning template guiding the model through: (1) whether the emotion can be clearly inferred from text alone, (2) whether audio or visual cues offer supplementary or clarifying information, and (3) whether multimodal signals contradict the textual content. This structure encourages explicit reasoning.
- **Context.** A multimodal context comprising the dialogue history, each utterance  $u_i$  is paired with its corresponding audio description  $C_{\text{aud}}^{(i)}$  and only the target utterance  $u_k$  is additionally accompanied by a visual description  $C_{\text{vid}}^{(k)}$ , and the underlined target utterance  $u_k$ .

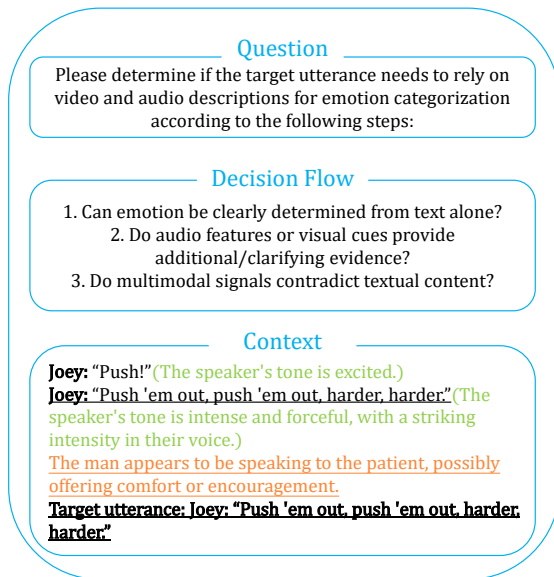


Figure 8: Question Prompt Example. The prompt consists of three components: an instruction (Question), a step-wise decision guide (Decision Flow), and a multimodal context (Context). Green text indicates audio descriptions  $C_{\text{aud}}^{(i)}$ , red text denotes the visual description  $C_{\text{vid}}^{(k)}$ , and the underlined sentence marks the target utterance  $u_k$ .

**Response Template** Figure 9 shows the chain-of-thought response produced by our model af-

ter receiving the question prompt: reasoning steps are enclosed within `<think>...</think>`, followed by a binary decision enclosed in `<answer>...</answer>`. The `<think>` block records the model's deductions according to the three checkpoints in the Decision Flow, while the `<answer>` block provides the final decision on whether to retain multimodal inputs. This structured output allows for consistent supervision and facilitates learning under reinforcement learning settings, where outputs that match the expected format are positively rewarded.

By chain-of-thought reasoning, enhances the interpretability and robustness of the multimodal selection process.

```
<think>1. The target utterance "Push 'em out, push 'em out, harder, harder" is emotionally clear and implies urgency and intensity.
2. The audio caption provides context but does not add significant emotional clarity since the text already conveys the emotion.
3. The image caption describes a medical setting but does not provide additional emotional context.</think>
<answer>No, the captions should be removed.</answer>
```

Figure 9: Response Example. The model produces structured reasoning steps within `<think>...</think>` followed by a binary decision in `<answer>...</answer>`. In this case, it determines that the multimodal features are redundant for emotion understanding.

**Emotion Recognition Prompt** Figure 10 presents the inference prompt template used for emotion classification. The user instruction defines the model as a sentiment expert, provides the conversation context and target utterance  $u_k$ , and conditionally appends audios  $C_{\text{aud}}$  and video description  $C_{\text{vid}}^{(k)}$  whenever the modality selection module indicates their necessity. The assistant must respond with a single emotion label  $e_i$ .

```
User
Now you are expert of sentiment and emotional analysis. The following conversation noted between...
Given the conversation and the characteristics of these speakers, {speaker utterance u_j}({speaker audio description C_aud})
Given the description of video, {video description C_vid}
Please select the emotional label of {utterance u_k} from <emotion labels>:

Assistant
{emotional label e_i}
```

Figure 10: Prompting template for emotion classification. Audio and visual descriptions (green) are included only if selected by the modality decision model.

Dataset	conversations			utterances			classes
	train	dev	test	train	dev	test	
IEMOCAP	108	12	31	5163	647	1623	6
MELD	1038	114	280	9989	1109	2610	7

Table 3: The statistics of datasets.

## A.2 Datasets

**MELD** contains 1,433 conversations and 13,708 utterances, each utterance is labeled with one of seven emotions: neutral, surprise, fear, sadness, joy, disgust, and anger.

**IEMOCAP** contains 153 conversations and 7,433 utterances. Each utterance is labeled with one of six emotions: happy, sad, neutral, angry, excited, and frustrated.

## A.3 Baselines

**DialogueRNN**(Majumder et al., 2019): It models speaker and sequential information in conversations through three different GRUs.

**MMGCN**(Hu et al., 2021): It utilizes the GCN network to obtain contextual information, which not only effectively exploits multimodal dependencies, but also makes full use of speaker information.

**DialogueTRM**(Mao et al., 2021): It extends the concept of emotion dynamics to multi-modal settings and proposes a transformer module for simultaneously modeling the intra-modal and inter-modal emotion dynamics.

**MM-DFN**(Hu et al., 2022a): It designs a graph-based dynamic fusion module to fuse multimodal contextual features, which reduces redundancy and enhances complementarity between modalities.

**EmoCaps**(Li et al., 2022): It proposes a new structure, Emoformer, to extract multimodal emotion vectors from different modalities and fuses them with sentence vectors to be an emotion capsule.

**MPT-HCL**(Zou et al., 2023): It designs a Multimodal Prompt Transformer (MPT) to perform cross-modal information fusion and uses the Hybrid Contrastive Learning (HCL) strategy to optimize the model’s ability to handle labels with few samples.

**TelME**(Yun et al., 2024): It incorporates cross-modal knowledge distillation to transfer information from a language model acting as the teacher to the non-verbal students, thereby optimizing the efficacy of the weak modalities and

then combine multimodal features using a shifting fusion approach.

**InstructERC**(Lei et al., 2023a): It introduces a simple but effective retrieval template module and two additional emotion alignment tasks to implicitly model the dialogue role relationships and future emotional tendencies in conversations.

**BiosERC**(Xue et al., 2024): It extract the “biographical information” of the speaker within a conversation as supplementary knowledge injected into the model to classify emotional labels for each utterance.

**VoiceERC**(Wu et al., 2025): It translates speech characteristics into natural language descriptions, allowing LLMs to perform multimodal emotion analysis via text prompts without any architectural changes.

Hyperparameter	Value
Batch	8
Epoch	3
First Stage Learning Rate	5e-6
Second Stage Learning Rate	5e-5
LoRA r	8
LoRA alpha	16
LoRA dropout	0.1
Group Generations	4

Table 4: Training hyperparameters used for both the GRPO-guided modality-selection and LoRA-tuned emotion-recognition stages.

## A.4 Experiment Setup

During the GRPO-guided Modality Selection stage, we adopt Qwen2.5-7B (Yang et al., 2024b) as the backbone model. In the subsequent LoRA-tuned LLMs for Emotion Recognition stage, all models, including our method and LLM baselines, are implemented with the LLaMA2-7B (Touvron et al., 2023) to ensure fair comparisons across experimental settings. To curb the number of trainable

















Modality	Utterance	True Label	w/o GRPO & modality desc.	w/o GRPO	Ours
 	U3: Push 'em out, push 'em out, way out!	Joy	Joy	Joy	Joy
 	U4: Let's get that ball and really move, hey, hey, ho, ho.	Joy	Joy	Joy	Joy
 	U5: Let's I was just yeah, right.	Joy	Neutral	Joy	Joy
 	U6: Push!	Joy	Anger	Joy	Joy
Modality	Utterance	True Label	w/o GRPO & modality desc.	w/o GRPO	Ours
 	U2: Umm, well I would have to say that it's a, it's tragic love story.	Neutral	Neutral	Neutral	Neutral
 	U3: Well, that's sort of a given, but yes. Anyone else?	Neutral	Neutral	Neutral	Neutral
 	U4: Oh-oh-oh, symbolism! And uh, the-the uh, wildness of the mores, which	Joy	Joy	Neutral	Joy
 	U5: Excellent! What Rachel has shrewdly observed here	Joy	Joy	Joy	Joy

Figure 11: case study

parameters while minimizing performance degradation, we apply LoRA. Our models are trained on two RTX 4090 GPUs (24 GB each) due to computational constraints. The remaining hyperparameter configurations are summarized in Table 4. To ensure robustness, we averaged the results from three independent runs.