# Beyond Modality Collapse: Taming Guided Modality Entropy for Omni-modal Emotion Reasoning

**Xian Zhao**[1,2], **Rui Hu**[1,2], **Yuxiang Zhang**[1], **Delai Qiu**[2], **Yining Wang**[2], **Shengping Liu**[2], **Jian Yu**[1], **Jitao Sang**[1,3*]

[1]Beijing Key Laboratory of Traffic Data Mining and Embodied Intelligence, Beijing Jiaotong University

[2]Unisound AI Technology Co., Ltd.

[3] State Key Laboratory of AI Safety, Beijing

{xian.zhao, jtsang}@bjtu.edu.cn

## Abstract

Omni-modal Large Language Models (OLLMs) excel in diverse tasks but struggle with complex emotional reasoning, which requires integrating textual, visual, and acoustic signals. We attribute this limitation to *modality collapse*, where models over-rely on a dominant modality while neglecting complementary cues. To address this issue, we introduce OmniCoT, a data paradigm that interleaves guided tokens (e.g., [vision], [audio]) into reasoning traces to enforce structured evidence extraction. To further internalize the reasoning behaviors instilled by OmniCoT and facilitate adaptive modality prioritization, we propose Dynamic Modality-Entropy GRPO (DyME-GRPO), which utilizes entropy-based uncertainty estimates over Guided Tokens (GTs) to regulate modality usage, thereby mitigating collapse and informational redundancy. By applying supervised fine-tuning with OmniCoT followed by DyME-GRPO, we develop EmoOmni based on the Qwen2.5-Omni-7B backbone. Extensive experiments demonstrate that EmoOmni achieves state-of-the-art performance on multiple emotion recognition and reasoning benchmarks while preserving the general capabilities of the base model. These findings highlight the potential of our work for omni-modal reasoning across a broader range of complex tasks.[1]

## 1 Introduction

As a pivotal challenge within the field of Human-centered Artificial Intelligence (Tocchetti et al., 2025), omni-modal emotion recognition and reasoning (Lian et al., 2023b; Cheng et al., 2024) requires the effective integration of textual, acoustic, and visual information. While Omni-modal Large Language Models (OLLMs) have demonstrated exceptional perceptual capabilities across

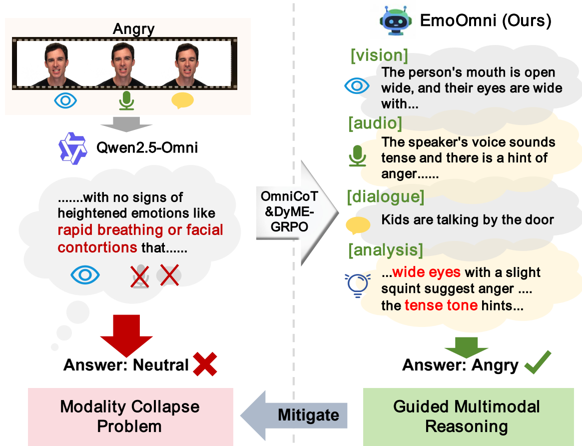



Figure 1: **Comparison of reasoning schemes for omni-modal emotion recognition.** While vanilla Qwen2.5-Omni incorrectly predicts "Neutral" state due to visual over-reliance, EmoOmni employs a token-guided CoT to integrate multimodal evidence, which effectively mitigates modality collapse and enable correct "Angry" identification.

a broad spectrum of multimodal tasks (Xu et al., 2025b; Yue et al., 2024), they still struggle with the complex emotional reasoning. We observe that this limitation stems primarily from *modality collapse*, a phenomenon where the model exhibits a disproportionate reliance on the dominant modality. As illustrated in Figure 1, vanilla Qwen2.5-Omni-7B (Xu et al., 2025a) tends to rely disproportionately on visual cues. Such visual-centric bias leads to the systematic marginalization of critical acoustic and linguistic evidence, thereby undermining the overall reliability and accuracy of the emotional reasoning process.

Motivated by the success of chain-of-thought (CoT) prompting in vision–language reasoning (Shao et al., 2024a; Shi et al., 2025; Man et al., 2025; Lin et al., 2025), which encourages models to decompose complex problems into step-by-step reasoning processes, we propose **OmniCoT**, a data construction paradigm designed to generate

[1]Our code is released at https://github.com/ZxShane/EmoOmni.



*Findings of the Association for Computational Linguistics: ACL 2026*, pages 20830–20845

structured and guided multimodal reasoning traces to mitigate modality collapse. OmniCoT incorporates **Guided Tokens (GTs)**, such as `[vision]`, `[audio]`, `[dialogue]`, and `[analysis]`, which serve as explicit cognitive anchors to instruct the model when to consult specific modalities or execute integrative synthesis, thereby encouraging modality-aware reasoning rather than uncontrolled free-form explanation. By leveraging the advanced logical capabilities of Large Language Models (LLMs), An LLM-based verification and integration module further guarantee the logical coherence and correctness of the generated CoT.

To further internalize the reasoning behaviors instilled by OmniCoT, we propose **Dy**namic **M**odality-**E**ntropy GRPO (**DyME-GRPO**), extending the Group Relative Policy Optimization (GRPO) (Shao et al., 2024b) to the multimodal domain. By quantifying predictive uncertainty through entropy-based estimates of modality GTs, this method dynamically penalizes high-uncertainty distributions to adaptively calibrate modality reliance. This facilitates a transition from static format imitation to dynamic modality coordination, allowing the model to converge toward a balanced policy that simultaneously mitigates modality collapse and suppresses informational redundancy.

Leveraging Qwen2.5-Omni-7B as a backbone, we develop EmoOmni through the integrated application of OmniCoT and DyME-GRPO. Experimental results demonstrate that EmoOmni achieves state-of-the-art performance across multiple benchmarks, significantly outperforming both vanilla and SFT-adapted baselines fine-tuned on the same emotion dataset, while preserving general multimodal capabilities. These findings highlight the potential of our work as a scalable paradigm for omni-modal reasoning across a broader spectrum of complex tasks.

Our contributions are as follows:

- We propose OmniCoT, a data construction paradigm that utilizes guided tokens and verification mechanisms to ensure the logical coherence of emotional inference.
- We introduce DyME-GRPO, an RL-based optimization strategy that adaptively calibrates modality reliance via entropy-based uncertainty, effectively mitigating modality collapse and avoiding informational redundancy.
- We develop EmoOmni, which achieves state-of-the-art performance on multiple multimodal emotion recognition and reasoning benchmarks and maintains the general multimodal capabilities.

## 2 Related Work

### 2.1 Omni-modal Emotion Recognition and Reasoning Datasets.

Omni-modal emotion recognition and reasoning aims to infer emotion by synthesizing video, audio, and textual signals, of which advancement relies heavily on datasets that provide both emotional labels and reasoning traces. While existing benchmarks such as EMER (Lian et al., 2023b) and MERR (Cheng et al., 2024) provide reasoning traces, and subsequent works (Lian et al., 2024; Yang et al., 2025b) have refined the inference pipeline, these datasets often rely on unstructured, free-form explanations that lack rigorous logical verification. To overcome these limitations, we propose OmniCoT, which enforces modality-aware reasoning via structured GTs and ensures logical reliability through an LLM-based verification module.

### 2.2 Omni-modal Large Language Models

OLLMs integrate diverse modalities into a unified processing framework. Following early proprietary models like GPT-4o (Hurst et al., 2024) and Gemini 1.5 (Team et al., 2024), open-source architectures such as VITA-1.5 (Fu et al., 2025) and EMOVA (Chen et al., 2025) successfully aligned multimodal encoders with LLM backbones. While foundational models such as MiniCPM-o (Yao et al., 2024) and Qwen2.5-Omni (Xu et al., 2025a) emphasize broad generalization, specialized variants like HumanOmni (Zhao et al., 2025b) and the reasoning R1-Omni (Zhao et al., 2025a) have begun to prioritize human-centric perception and reinforcement learning for emotional reasoning. Most recently, the Qwen3-Omni series (Xu et al., 2025b) has advanced OLLM through MoE architectures and reasoning processes. We evaluate our proposed **EmoOmni** against these representative baselines to demonstrate our superiority.

## 3 OmniCoT: Token-Guided Multimodal CoT Construction

### 3.1 Overview

OmniCoT introduces a token-guided reasoning paradigm to mitigate modality collapse in OLLMs

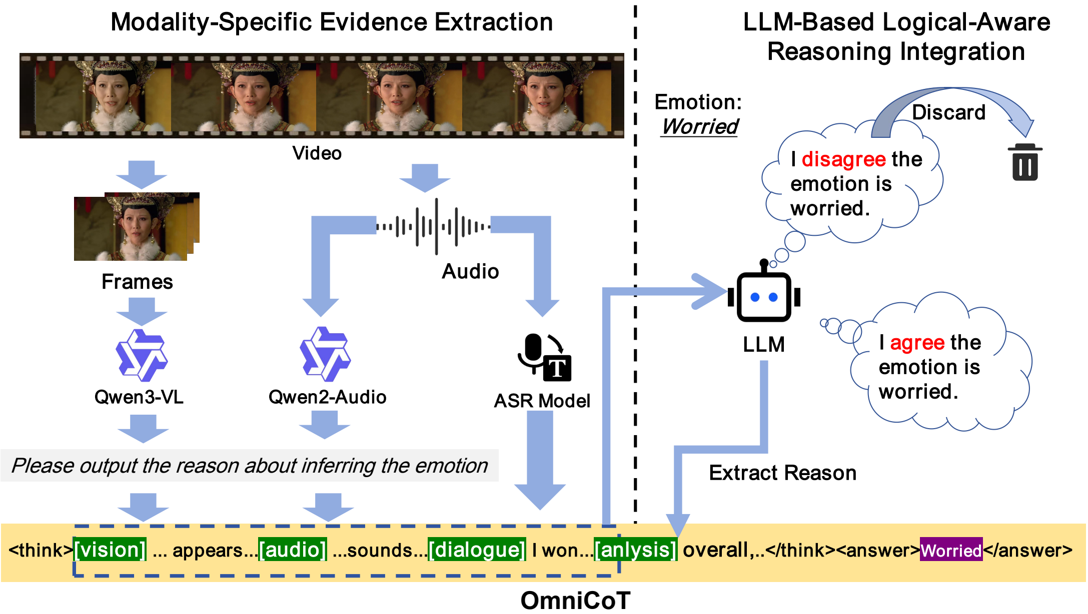



Figure 2: Overview of the OmniCoT. Left: **Modality-Specific Evidence Extraction**: Modality-specific evidence is extracted via MLLMs. Rgith: **LLM-Based Logical-Aware Reasoning Integration**: Multimodal CoT are filtered by an LLM-based verification module for logical coherence.

by enforcing the coordination of multimodal signals through structured Guided Tokens (GTs). As illustrated in Figure 2, modality-specific tokens including `[vision]`, `[audio]`, and `[dialogue]` systematically invoke distinct evidence while the `[analysis]` token initiates a deliberate reasoning phase. To ensure logical coherence, an LLM-based verification module refines these reasoning traces. By formalizing modality utilization, OmniCoT enables OLLMs to internalize modality-aware behaviors and eliminates the over-reliance on a dominant modality.

### 3.2 Modality-Specific Evidence Extraction

The pipeline begins with the extraction of fine-grained, modality-specific evidence (Figure 2, left). For visual modality, Qwen3-VL (Bai et al., 2025) derives emotion-relevant descriptions including postures, expressions, and environmental cues for appending to the `[vision]` token. For audio modality, Qwen2-Audio (Chu et al., 2024) captures paralinguistic attributes such as intonation and vocal intensity following the `[audio]` token. Simultaneously, Whisper (Radford et al., 2023) transcribes spoken dialogue for the `[dialogue]` token. Detailed prompts for these models are provided in Appendix A. To ensure factual and logical integrity, an LLM-based verification filters the extracted evidence before its final synthesis into the CoT.

### 3.3 LLM-Based Logical-Aware Reasoning Integration

Simply concatenating fragmented multimodal information may result in information insufficiency or logical inconsistencies. To resolve this, we leverage the advanced logical capabilities of LLMs by employing Qwen3-7B (Yang et al., 2025a) as a core integration and verification module (Figure 2, right). This module implements a rigorous consistency verification whereby the LLM assesses whether the multimodal evidence provides sufficient support for the ground-truth emotion. Upon successful validation, the LLM generates a definitive reasoning step appended to the `[analysis]` token to finalize the OmniCoT structure, while the failure examples are excluded from the dataset. This verification mechanism ensures only semantically aligned and logically robust examples are retained, thereby safeguarding the overall quality of the synthesized CoT data. Detailed prompts are provided in Appendix A.

## 4 EmoOmni: Calibrating Omni-modal Emotion Reasoning via OmniCoT and DyME-GRPO

### 4.1 Overview

EmoOmni is built upon Qwen2.5-Omni-7B (Xu et al., 2025a), leveraging its native architecture

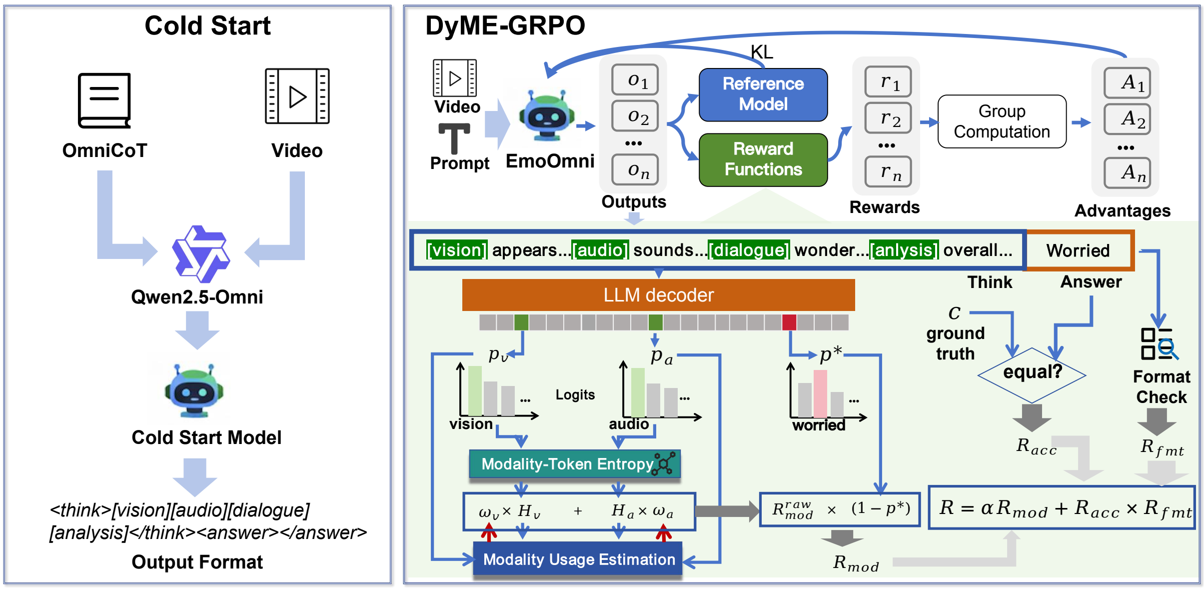



Figure 3: Overview of EmoOmni training pipeline. Left: **Cold Start** via SFT on OmniCoT to internalize structured reasoning traces. Right: **DyME-GRPO** reinforcement learning framework, which optimizes the reasoning policy by incorporating a dynamic modality-entropy reward ($R_{mod}$) to mitigate modality collapse alongside standard accuracy ($R_{acc}$) and format ($R_{fmt}$) rewards.

for processing interleaved multimodal signals. To optimize the reasoning policy, we employ Group Relative Policy Optimization (GRPO) (Shao et al., 2024b), which refines model performance by evaluating sampled outputs against relative rewards within a group. A more detailed introduction to the GRPO framework is provided in Appendix B. Our training follows a two-stage pipeline where the model first undergoes Supervised Fine-Tuning (SFT) on the OmniCoT dataset to internalize the structured, token-guided reasoning format. Building upon this initialized policy, we introduce DyME-GRPO, which incorporates a dynamic modality-entropy reward to adaptively calibrate modality reliance through entropy-based signals. This transition from static format imitation to dynamic modality coordination empowers the model to alleviate modality collapse while effectively mitigating informational redundancy.

## 4.2 Cold start

To align the model with the GTs and structured reasoning defined in OmniCoT, we first perform SFT on the OmniCoT dataset as a cold start stage. Given a multimodal prompt $x$ and its corresponding OmniCoT sequence $y$, the SFT objective minimizes the negative log-likelihood of the target sequence:

$$\mathcal{L}_{\text{SFT}} = -\sum_{t} \log P_\theta(y_t \mid y_{<t}, x)\,, \qquad (1)$$

where $y_t$ denotes the $t$-th token of the target reasoning chain. This stage ensures the model can correctly invoke GTs and adhere to the prescribed CoT format, providing a stable policy initialization for subsequent DyME-GRPO stage.

## 4.3 DyME-GRPO

Building on the SFT policy, we employ DyME-GRPO to refine multimodal reasoning. As shown in the bottom of Figure 3, the total reward $R$ comprises three components:1) **Format reward** ($R_{fmt}$) penalizes deviations from the Omni-CoT template and GTs usage. 2) **Accuracy reward** ($R_{acc}$) evaluates the correctness of the final emotional inference. 3) **Dynamic modality-entropy reward** ($R_{mod}$) adaptively calibrates modality reliance using entropy-based signals to penalize informational redundancy and reward discriminative cues.

### 4.3.1 Format reward

The format reward encourages the model to produce responses that follow the desired token-guided CoT structure, including the correct use of GTs such as `[vision]`, `[audio]`, `[dialogue]`, and `[analysis]`. Let $y$ denote the generated output, the format reward is defined as:

$$R_{\text{fmt}}(y) = \begin{cases} 1, & \text{if } y \text{ in OmniCoT format} \\ 0, & \text{otherwise.} \end{cases} \qquad (2)$$

#### 4.3.2 Accuracy reward

The accuracy reward aligns the optimization objective with the target emotion recognition task. Let $c$ denote the ground-truth emotion label and $\hat{c} = \Phi(y)$ be the predicted label extracted from the reasoning chain $y$. We define this reward as:

$$R_{\text{acc}} = \mathbb{I}[\hat{c} = c], \tag{3}$$

where $\mathbb{I}[\cdot]$ is the indicator function that returns 1 when the prediction is correct and 0 otherwise.

#### 4.3.3 Dynamic Modality-Entropy Reward

**Modality-Token Entropy:** To quantify the uncertainty in modality selection, we evaluate the entropy at the decision points where modality-specific tokens (e.g., `[vision]`, `[audio]`) are generated. Let $t_m$ denote the position of a modality token $m \in \{\text{vision, audio}\}$. The model's probability distribution over the vocabulary $\mathcal{V}$ at $t_m$ is denoted as $p_{t_m,k}$ for token $k \in \mathcal{V}$. We define the modality entropy $H_m$ as the Shannon entropy (Shannon, 1948) of this distribution:

$$H_m = -\sum_{k\in\mathcal{V}} p_{t_m,k} \log p_{t_m,k}. \tag{4}$$

$H_m$ serves as a proxy for modality-level ambiguity; higher entropy reflects increased uncertainty or exploratory behavior during modality selection.

**Modality Usage Estimation:** We estimate the model's reliance on each modality by integrating both selection confidence and uncertainty. Let $p_m$ be the softmax probability of the selected modality token $m$ at position $t_m$. We define the unnormalized usage intensity $u_m$ as:

$$u_m = \frac{p_m}{H_m}, \tag{5}$$

where $u_m$ increases when a modality is selected with high confidence (large $p_m$) and low uncertainty (small $H_m$). Let $v$ and $a$ denote vision and audio modality. To facilitate cross-modal comparison, these intensities are normalized into relative weights:

$$w_m = \frac{u_m}{\sum_{j\in\{\text{v, a}\}} u_j}. \tag{6}$$

The resulting weights $w_m \in [0, 1]$ satisfy $\sum w_m = 1$, providing a continuous measure of the model's relative reliance on each modality. A higher $w_m$ indicates a more decisive and prioritized utilization of modality $m$ within the reasoning chain.

**Reward function:** We integrate these signals into the reinforcement learning objective through a usage-weighted entropy term:

$$R^{\text{raw}}_{\text{mod}} = w_{\text{v}}\, H_{\text{v}} + w_{\text{a}}\, H_{\text{a}}. \tag{7}$$

To adaptively modulate exploratory behavior, we scale this term by the model's uncertainty regarding the ground-truth label. Let $p^*$ denote the predicted probability of the correct answer. The final **Dynamic Modality-Entropy Reward** is defined as:

$$R_{\text{mod}} = (1 - p^*) R^{\text{raw}}_{\text{mod}}. \tag{8}$$

The factor $1 - p^*$ acts as a dynamic scaling factor that suppresses exploration as the model gains confidence in the correct inference ($p^* \to 1$). Conversely, lower confidence encourages the integration of complementary multimodal cues to mitigate over-reliance on any single source and foster robust, modality-aware reasoning.

#### 4.3.4 Overall reward function

We combine the above components into a unified reinforcement learning objective:

$$R = R_{\text{fmt}} \cdot R_{\text{acc}} + \alpha\, R_{\text{mod}}, \tag{9}$$

where $\alpha > 0$ is a balancing coefficient, we set $\alpha = 0.1$ in emoOmni. More discussion about $\alpha$ can be found in ablation study. The product of $R_{\text{fmt}}$ and $R_{\text{acc}}$ ensures that accuracy is only rewarded when the correct format is maintained. Meanwhile, $R_{\text{mod}}$ encourages the model to explore different modalities instead of just guessing the answer. Together, these rewards guide the model to produce reasoning that is both well-structured and accurate.

## 5 Experiments

### 5.1 Settings

**Training Configurations:** Building on the Qwen2.5-Omni-7B backbone, we obtain the EmoOmni model via the two-stage training procedure detailed in Section 4. For the cold start stage, we utilize 4.4K OmniCoT examples sampled from the MER 2023 dataset (Lian et al., 2023a). The subsequent GRPO stage employs 17K training examples derived from DFEW (Jiang et al., 2020) and the training set of MER 2025 datasets (Lian et al., 2025).

**Baselines and Comparison:** We evaluate EmoOmni against three categories of representative models: 1) vanilla OLLMs, 2) OLLMs fine-tuned on the same emotion datasets without CoT to

Table 1: Main results on multimodal emotion recognition and reasoning benchmarks. We report WA (Weighted Accuracy), UAR (Unweighted Average Recall), and WAR (Weighted Average Recall) cross four benchmarks, where **bold** and underlined values denote the best and second-best performance, respectively. EmoOmni achieves state-of-the-art results across most metrics and demonstrates competitive performance on the remainder.

| Model | Variant | RAVDESS | | | MELD | | | IEMOCAP | | | MER2025 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | WA | UAR | WAR | WA | UAR | WAR | WA | UAR | WAR | WA | UAR | WAR |
| *vanilla Omni-modal large language models* | | | | | | | | | | | | | |
| HumanOmni | vanilla | 48.79 | 55.61 | 55.64 | 42.86 | 30.65 | 42.64 | 65.03 | 36.00 | 43.97 | 68.77 | 60.22 | 72.16 |
| VITA-1.5 | vanilla | 51.85 | 32.81 | 37.98 | 38.56 | 25.74 | 31.25 | 56.15 | 28.12 | 42.54 | 64.42 | 45.48 | 49.28 |
| MiniCPM-o 2.6 | vanilla | 48.79 | 55.64 | 55.61 | 8.24 | 11.11 | 11.76 | 12.32 | 10.35 | 13.72 | 26.13 | 34.62 | 38.14 |
| Qwen2.5-Omni-7B | vanilla | 68.26 | 51.01 | 53.54 | 44.65 | 35.42 | 46.12 | 62.55 | 41.33 | 52.92 | 65.63 | 62.47 | 67.46 |
| Qwen2.5-Omni-3B | vanilla | 46.78 | 43.17 | 45.75 | 47.51 | 31.08 | **49.71** | 61.64 | 29.70 | 37.72 | 67.24 | 40.18 | 52.27 |
| Qwen3-Omni-30B-Instruct | vanilla | 63.76 | 52.95 | 54.81 | 47.39 | 26.55 | 34.88 | 64.65 | 42.40 | 47.53 | 72.88 | 68.65 | 69.02 |
| *Omni-modal large language models with SFT* | | | | | | | | | | | | | |
| HumanOmni | SFT | 46.73 | 51.12 | 52.24 | 34.36 | 24.98 | 32.81 | 62.97 | 37.06 | 46.82 | 73.05 | 66.70 | **77.49** |
| VITA-1.5 | SFT | 47.07 | 45.17 | 47.68 | 39.07 | 23.68 | 41.86 | 60.81 | 29.08 | 41.79 | 69.86 | 63.30 | 64.73 |
| MiniCPM-o 2.6 | SFT | 11.97 | 17.04 | 17.87 | 25.55 | 15.64 | 32.56 | 18.75 | 15.26 | 17.40 | 28.32 | 33.27 | 39.05 |
| Qwen2.5-Omni-7B | SFT | 64.65 | 55.61 | 55.85 | 40.40 | 30.37 | 38.76 | 65.83 | 40.90 | 52.73 | 70.06 | 60.33 | 64.31 |
| Qwen2.5-Omni-3B | SFT | 60.27 | 50.14 | 50.64 | 42.27 | 28.91 | 43.74 | 60.42 | 31.42 | 40.41 | 71.91 | 56.45 | 66.54 |
| Qwen3-Omni-30B-Instruct | SFT | 61.85 | 52.83 | 51.32 | 42.33 | 30.31 | 37.98 | 63.82 | 33.82 | 38.17 | **75.92** | 70.43 | 76.44 |
| *Omni-modal reasoning models* | | | | | | | | | | | | | |
| R1-Omni | vanilla | 49.24 | 56.04 | 56.12 | 33.94 | 27.83 | 31.78 | 43.49 | 27.74 | 41.74 | 49.71 | 37.11 | 42.44 |
| Qwen3-Omni-30B-Think | vanilla | 65.14 | 55.30 | 55.69 | 50.23 | 36.50 | 35.94 | 58.30 | 25.91 | 44.75 | 63.68 | 27.97 | 30.70 |
| **EmoOmni(ours)** | / | **70.28** | **62.33** | **62.34** | **48.60** | **36.64** | 46.53 | **67.12** | **42.89** | **58.15** | 73.94 | **71.59** | 75.42 |

ensure a rigorous and fair comparison and 3) Omni-modal reasoning models. This setup allows us to isolate the impact of structured reasoning from simple data exposure.

**Evaluation Benchmarks and Metrics:** Models are assessed across four widely-used multimodal emotion benchmarks. Among these, RAVDESS (Livingstone and Russo, 2018) and IEMOCAP (Busso et al., 2008) consist of acted emotional expressions, while MELD (Poria et al., 2019) and MER 2025 collect emotion-rich clips sourced from films. Performance is quantified using standard metrics such as Weighted Accuracy (WA), Unweighted Average Recall (UAR) and Weighted Average Recall (WAR).

Implementation details about experiments are deferred to Appendix C.

## 5.2 Quantitative Results on Emotion Benchmarks

### 5.2.1 Comparison with Omni-modal Large Language Models

Compared with vanilla OLLMs, EmoOmni achieves superior results across nearly all evaluation metrics, with competitive performance on the remainder. Such consistency across diverse benchmarks suggests that EmoOmni derives its robustness from generalized reasoning rather than from dataset-specific biases.

Table 2: Impact of Guided Tokens (GTs) on model performance. OmniCoT outperforms standard CoT variants. The performance drop in the w/o GTs setting confirms that explicit GTs are indispensable for coherent multimodal reasoning.

| Dataset | CoT | GTs | MELD | | IEMOCAP | |
|---|---|---|---|---|---|---|
| | | | UAR | WAR | UAR | WAR |
| MER2023 | ✗ | ✗ | 28.54 | 38.57 | 38.92 | 50.87 |
| MERR | ✓ | ✗ | 27.96 | 43.06 | 41.70 | 55.89 |
| MER2025 | ✓ | ✗ | 29.78 | 38.56 | 36.47 | 51.08 |
| OmniCoT | ✓ | ✓ | **34.30** | **42.13** | **44.63** | **57.75** |
| *w/o GTs* | ✓ | ✗ | 33.80 | 44.09 | 40.74 | 56.87 |

To isolate the impact of our methodology, we fine-tuned OLLM baselines on the same data source used for EmoOmni. While this controlled SFT significantly enhances performance on the in-domain MER2025 benchmark, its gains on other datasets are marginal. This indicates that vanilla SFT primarily facilitates domain adaptation but fails to cultivate the cross-dataset generalization essential for robust emotional inference. Despite the competitive advantage of equivalent data exposure, SFT-adapted OLLMs still trail EmoOmni. This persistent performance gap underscores that EmoOmni's superiority stems from the structural innovations of OmniCoT and DyME-GRPO rather

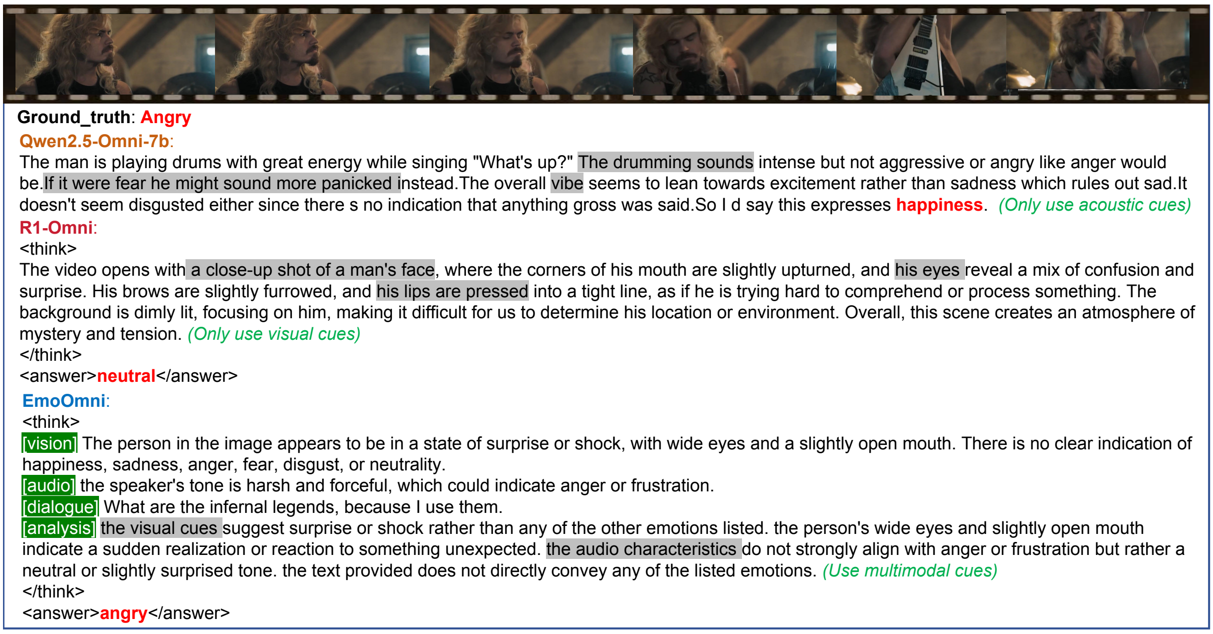



Figure 4: Qualitative comparison of reasoning processes between EmoOmni and representative OLLMs. EmoOmni utilizes structured guided tokens (`[vision]`, `[audio]`, `[dialogue]`) to explicitly extract evidence. During the `[analysis]` phase, the model achieve an accurate inference by grounding its decision in specific acoustic and visual cues.

than mere data volume.

#### 5.2.2 Compared with Omni-modal Reasoning Models

EmoOmni consistently outperforms specialized reasoning models, including R1-Omni and the substantially larger Qwen3-Omni-30B-Thinking. Although R1-Omni pioneered GRPO-based emotional reasoning, its vanilla objective lacks specialized multimodal optimization. In contrast, EmoOmni outperforms across all benchmarks, demonstrating that DyME-GRPO more effectively advances multimodal emotional inference. This superiority confirms that principled modality coordination is essential for mastering complex affective tasks.

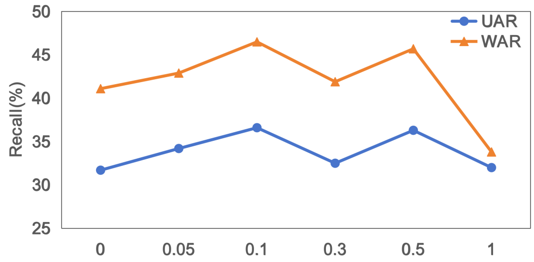
### 5.3 Qualitative Analysis

Figure 4 compares the reasoning processes of Qwen2.5-Omni, R1-Omni, and EmoOmni. While Qwen2.5-Omni often relies on single-modality cues and R1-Omni generates reasoning steps lacking clear modality grounding, EmoOmni follows a structured, modality-aware path. By explicitly identifying evidence through `[vision]`, `[audio]`, and `[dialogue]` tokens before the final `[analysis]`, EmoOmni resolves complex cases where individual modalities might be misleading. This systematic integration confirms that our guided tokens produce more coherent and grounded multimodal predictions, with additional examples provided in Appendix D.

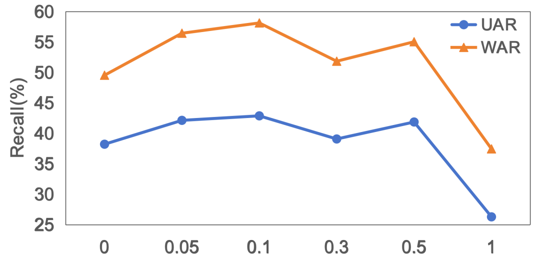
### 5.4 The Necessity of Guided Tokens

Existing datasets like MERR and MER 2025 provide multimodal chains of thought but lack explicit GTs to direct the model across modalities. To evaluate the impact of these tokens, we compare Qwen2.5-Omni fine-tuned on: (i) non-CoT data, (ii) existing CoT datasets, and (iii) our OmniCoT. We also include a variant of OmniCoT where the reasoning chains are preserved but the GTs are removed. All training sets are identical in size to ensure a fair comparison. Examples of each CoT are illustrated in Appendix E.

As shown in Table 2, while standard CoT fine-tuning improves performance, Omni-CoT achieves the best results. Removing GTs leads to a noticeable performance drop, confirming that explicit GTs are essential for structured multimodal reasoning. These findings demonstrate that the efficacy of OmniCoT stems not only from the reasoning text itself, but also from the structural discipline imposed by GTs, which serve as the key for superior emotional reasoning.




(a) Results on MELD dataset




(b) Results on IEMOCAP dataset

Figure 5: Impact of $\alpha$ on performance. The results show that Dynamic Modality-Entropy reward improves performance over the $\alpha = 0$ baseline, with an optimal peak at $\alpha = 0.1$. This highlights a successful trade-off between multimodal exploration and output stability.

### 5.5 Ablation Studies

#### 5.5.1 The Effectiveness of DyME-GRPO

We evaluate DyME-GRPO by comparing EmoOmni with its SFT-only counterpart (Table 1 vs. Table 2). On MELD, GRPO increases UAR from 34.30 to 36.64 and WAR from 42.13 to 46.53. On IEMOCAP, while WAR rises to 58.15, the UAR slightly shifts to 42.89 due to a strategic trade-off where the model mitigates majority-class bias to favor precise discrimination in minority categories (see Appendix F). These results demonstrate that DyME-GRPO goes beyond simple accuracy boosting by facilitating more balanced predictions across diverse emotions, which confirms its critical role in enhancing cross-dataset generalization beyond the limits of standard supervised fine-tuning.

#### 5.5.2 Analysis of Hyper-parameter $\alpha$

The coefficient $\alpha$ in Eq. 9 balances the Dynamic Modality-Entropy reward against task-specific objectives. As shown in Figure 5b, performance on MELD and IEMOCAP follows a parabolic trend, peaking at $\alpha = 0.1$. Setting $\alpha = 0$ leads to a noticeable decline, confirming the necessity of entropy-driven regularization for effective modality integration. However, while a small $\alpha$ provides insufficient incentive, an excessively large value overemphasizes modality exploration at the expense of output stability and format adherence. Thus, $\alpha = 0.1$ is identified as the optimal trade-off between multimodal regularization and task-specific accuracy.

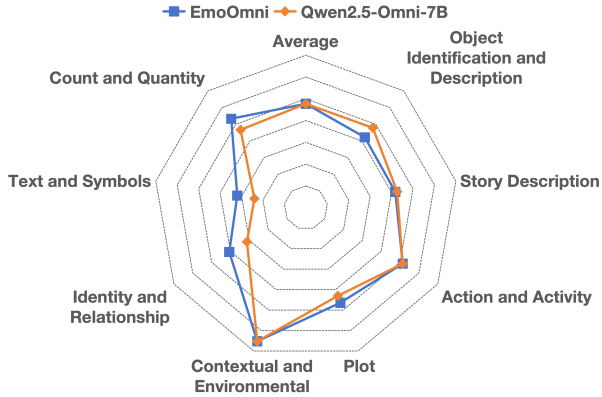



Figure 6: Generalization on OmniBench. EmoOmni preserves the base model's general capabilities without degradation.

### 5.6 Generalization on General Multimodal Tasks

To ensure that specializing in emotion reasoning does not compromise model versatility, we evaluate EmoOmni against the base Qwen2.5-Omni model on OmniBench (Li et al., 2024), covering eight multimodal categories. As illustrated in Figure 6, EmoOmni achieves performance comparable to or slightly exceeding the base model across all tasks, demonstrating its immunity to catastrophic forgetting. These results confirm that EmoOmni significantly advances emotional reasoning while fully preserving the general multimodal capabilities of the underlying OLLM.

## 6 Conclusion

We observe the modality collapse in OLLMs for emotion reasoning. To address this issue, we propose OmniCoT to enforce structured evidence extraction via modality-specific tokens and DyME-GRPO to adaptively balance modality reliance using entropy-based uncertainty. Our model, EmoOmni, outperforms representative OLLMs and Omni-multimodal reasoning models on most metrics while preserving general capabilities. These findings demonstrate the efficacy of structured reasoning and dynamic optimization, suggesting great potential for this framework to enhance a wide range of complex multimodal reasoning tasks beyond the emotion reasoning.

## Limitations

While EmoOmni achieves significant advancements, its scope remains subject to several constraints. First, this investigation is primarily centered on the multimodal emotion reasoning domain. Although our results on OmniBench demonstrate that general capabilities are preserved, the effectiveness and transferability of the DyME-GRPO framework to other specialized reasoning tasks, require further empirical validation. Second, EmoOmni is a generalized framework focusing on balancing modality usage, but it lacks specific designs for handling severe inter-modal contradictions. In cases where visual and acoustic cues convey conflicting emotions, our current entropy-driven method may need additional fine-grained strategies to resolve these ambiguities effectively.

## 7 Acknowledgment

This work is supported by the National Key RD Program of China (No. 2023YFC3310700) , Open Funding Programs of State Key Laboratory of AI Safety, and the National Natural Science Foundation of China (No. 2576030).

## References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

Kai Chen, Yunhao Gou, Runhui Huang, Zhili Liu, Daxin Tan, Jing Xu, Chunwei Wang, Yi Zhu, Yihan Zeng, Kuo Yang, and 1 others. 2025. Emova: Empowering language models to see, hear and speak with vivid emotions. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5455–5466.

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

Yunfei Chu, Jin Xu, Qian Yang, Wei Wei, Chenxie Wei, Ziyue Leng, Xingjun Lv, Jingyong He, Junyang Zhu, Chang Zhou, and 1 others. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Chaoyou Fu, Haojia Lin, Xiong Wang, Yi-Fan Zhang, Yunhang Shen, Xiaoyu Liu, Haoyu Cao, Zuwei Long, Heting Gao, Ke Li, and 1 others. 2025. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. *arXiv preprint arXiv:2501.01957*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

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

Yizhi Li, Ge Zhang, Yinghao Ma, Ruibin Yuan, Kang Zhu, Hangyu Guo, Yiming Liang, Jiaheng Liu, Zekun Wang, Jian Yang, and 1 others. 2024. Omnibench: Towards the future of universal omni-language models. *arXiv preprint arXiv:2409.15272*.

Zheng Lian, Rui Liu, Kele Xu, Bin Liu, Xuefei Liu, Yazhou Zhang, Xin Liu, Yong Li, Zebang Cheng, Haolin Zuo, and 1 others. 2025. Mer 2025: When affective computing meets large language models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 13837–13842.

Zheng Lian, Haiyang Sun, Licai Sun, Kang Chen, Mingyu Xu, Kexin Wang, Ke Xu, Yu He, Ying Li, Jinming Zhao, and 1 others. 2023a. Mer 2023: Multi-label learning, modality robustness, and semi-supervised learning. In *Proceedings of the 31st ACM international conference on multimedia*, pages 9610–9614.

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

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

Zhiyu Lin, Yifei Gao, Xian Zhao, Yunfan Yang, and Jitao Sang. 2025. Mind with eyes: from language reasoning to multimodal reasoning. *arXiv preprint arXiv:2503.18071*.

Steven R Livingstone and Frank A Russo. 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.

Yunze Man, De-An Huang, Guilin Liu, Shiwei Sheng, Shilong Liu, Liang-Yan Gui, Jan Kautz, Yu-Xiong Wang, and Zhiding Yu. 2025. Argus: Vision-centric reasoning with grounded chain-of-thought. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14268–14280.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 527–536.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*. PMLR.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Claude E Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. 2024a. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37:8612–8642.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024b. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Yudi Shi, Shangzhe Di, Qirui Chen, and Weidi Xie. 2025. Enhancing video-llm reasoning via agent-of-thoughts distillation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8523–8533.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Andrea Tocchetti, Lorenzo Corti, Agathe Balayn, Mireia Yurrita, Philip Lippmann, Marco Brambilla, and Jie Yang. 2025. Ai robustness: a human-centered perspective on technological challenges and opportunities. *ACM Computing Surveys*, 57(6):1–38.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, and 1 others. 2025a. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, and 19 others. 2025b. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

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

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*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

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

Jiaxing Zhao, Qize Yang, Yixing Peng, Detao Bai, Shimin Yao, Boyuan Sun, Xiang Chen, Shenghao Fu, Xihan Wei, Liefeng Bo, and 1 others. 2025b. Humanomni: A large vision-speech language model for human-centric video understanding. *arXiv preprint arXiv:2501.15111*.

# A Prompts for OmniCoT Generation

In this section, we provide the detailed prompts used during the OmniCoT construction pipeline to ensure the generation of high-quality, modality-aware reasoning traces.

## A.1 Multimodal Evidence Extraction

As illustrated in Figure 7, we utilize specialized MLLMs (Qwen3-VL and Qwen2-Audio) to perform fine-grained evidence extraction from raw multimodal signals:

**Visual Modality**: The prompt instructs the model to describe emotion-relevant visual cues,

**The Prompts for MLLMs**

**System Prompt:** You are an emotion analysis expert. Please infer emotion label based on the given video/audio.
**User Prompt:** Please infer one emotion among nine emotions: neutral, happy, sad, angry, fearful, worried and surprised.
Please output your reason about inferring the emotion:
Reason:{{The reason why you choose the emotion}}

Figure 7: Prompts of MLLMs for multimodal information extraction

**The Prompts for LLM**

**System Prompt:** You are an emotion analysis expert. Please infer emotion label based on the given multimodal features.
**User Prompt:** There are vision, audio and text modality analysis about a video.
{{Multimodal CoT}}
Please judge if the emotion of the person is {{answer}} and output 'yes' or 'no'. if output 'yes', please explain why.

Figure 8: Prompts of LLMs for logical-aware verification.

which are subsequently anchored by the [vision] token.

**Acoustic Modality**: For audio signals, the prompt directs the model to capture paralinguistic features including intonation, speech rate, and vocal intensity, placed following the [audio] token.

### A.2 Logical-Aware Verification and Integration

To prevent logical contradictions or information insufficiency, an LLM-based verification module is employed, as shown in Figure 8:

**Consistency Check**: The LLM is prompted to evaluate whether the extracted multimodal evidence (vision, audio, and text) logically supports the ground-truth emotion label.

**Reasoning Synthesis**: If the evidence is validated, the LLM generates a cohesive reasoning path that synthesizes the diverse cues into a final analysis, triggered by the [analysis] token.

**Filtering Criterion**: Any examples exhibiting contradictory signals or insufficient evidence for a definitive emotional inference are discarded to maintain dataset rigor.

## B Background on Group Relative Policy Optimization (GRPO)

We outline the mechanics of Group Relative Policy Optimization (GRPO), which serves as the algorithmic foundation for our proposed **DyME-GRPO**.

### B.1 Mechanism and Advantages

Unlike traditional Actor-Critic reinforcement learning frameworks, such as PPO (Schulman et al., 2017), which require an additional Critic model to estimate the value function, GRPO derives the baseline directly from the relative performance of a group of sampled outputs. As shown in Figure 3, for a given prompt $q$, the model samples a group of $G$ outputs $\{o_1, o_2, \ldots, o_G\}$. By calculating the mean and standard deviation of rewards within this group, GRPO eliminates the need for an explicit Critic network. This significantly reduces computational overhead and VRAM consumption, enabling the optimization of large-scale omni-modal models like Qwen2.5-Omni.

### B.2 Objective Function

The GRPO objective maximizes a surrogate loss that encourages the policy $\pi_\theta$ to produce outputs with higher-than-average rewards while staying close to a reference policy $\pi_{ref}$. The objective function is defined as:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \frac{1}{G}\sum_{i=1}^{G}\Bigg[\min\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}\hat{A}_i,\right. \left. \text{clip}\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}, 1-\epsilon, 1+\epsilon\right)\hat{A}_i\right) -\beta\mathbb{D}_{\text{KL}}(\pi_\theta \parallel \pi_{\text{ref}})\Bigg] \tag{10}$$

where $\epsilon$ is the clipping parameter and $\beta$ is the KL divergence penalty coefficient. The relative advantage $\hat{A}_i$ is computed by normalizing the rewards $r_i$ within the sampled group:

$$\hat{A}_i = \frac{r_i - \text{mean}(\{r_1, \ldots, r_G\})}{\text{std}(\{r_1, \ldots, r_G\}) + \eta} \tag{11}$$

In our **DyME-GRPO** framework, the reward $r_i$ is instantiated as the composite signal $R = \alpha R_{\text{mod}} + R_{\text{acc}} \times R_{\text{fmt}}$, where $R_{\text{mod}}$ leverages the entropy of guided tokens to calibrate multimodal reliance.

# C Experiments Settings

## C.1 Training Configurations

We adopt a two-stage training strategy to transition EmoOmni from structured format imitation to dynamic modality coordination. The details of the training stages and the datasets employed are as follows:

### C.1.1 Stage 1: Cold-start SFT with OmniCoT

In the first stage, we focus on aligning the backbone Qwen2.5-Omni with our proposed token-guided reasoning paradigm.

**Dataset**: We utilize the MER 2023 dataset, a large-scale multimodal benchmark featuring emotion-rich video clips from Chinese movies and TV series. Its naturalistic human interactions and multi-label annotations provide

**Data Construction**: We sampled 4.4K samples from MER 2023 and transformed them into OmniCoT format. For each sample, we randomly select one prompt from a pre-defined instruction list to enhance the model's instruction-following robustness. The specific prompt list used for SFT is provided in Figure 9.

### C.1.2 Stage 2: DyME-GRPO Reinforcement Learning

In this stage, we transition the model from structured imitation to autonomous affective reasoning. By leveraging reinforcement learning, the model learns to internalize the reasoning process, moving beyond superficial pattern matching to a deeper understanding of multimodal emotional cues.

**Datasets (DFEW & MER 2025)**: We expanded the training pool to 17K examples derived from two benchmarks: DFEW, which features dynamic facial expressions in unconstrained environments, and the training set of MER 2025, which focuses on complex affective reasoning for OLLMs

**Objective**: We apply DyME-GRPO to calibrate modality reliance. The model is optimized using a unified reward function combining format, accuracy, and the novel Dynamic Modality-Entropy Reward.

## C.2 Training Implementation Detail

All experiments were conducted on a high-performance computing cluster equipped with 4×NVIDIA A100 (80GB) GPUs, utilizing the ms-swift library for memory-efficient training. Our model is built upon the Qwen2.5-Omni backbone and fine-tuned using the LoRA (Low-Rank Adaptation) technique. Specifically, we set the LoRA rank to $r = 8$ and the scaling factor to $\alpha = 32$. During the SFT stage, we employed a constant learning rate of $1 \times 10^{-5}$ with a global batch size of 32. For the subsequent DyME-GRPO stage, we maintained the same learning rate while setting the balancing coefficient $\alpha_{reward}$ for the modality-entropy reward to 0.1. This optimal value, determined through extensive ablation studies, ensures an ideal trade-off between active multimodal exploration and output stability.

## C.3 Evaluation Configurations

### C.3.1 Evaluation Benchmarks

We assess our model on four widely-recognized benchmarks that capture different aspects of emotional expression:

**Acted Emotional Expression**: We use RAVDESS and IEMOCAP, which consist of high-quality recordings of professional actors performing specific emotional states.

**Naturalistic Emotion (In-the-wild)**: We evaluate on MELD and MER 2025, which contain emotion-rich clips sourced from films and TV series, offering more complex and spontaneous emotional scenarios.

### C.3.2 Evaluation Metrics

Performance is quantified using Weighted Accuracy (WA), Unweighted Average Recall (UAR), and Weighted Average Recall (WAR). Let $C$ be the number of emotion categories, $N$ be the total number of samples, and $N_i$ be the number of samples in class $i$. Let $n_{ii}$ denote the number of samples in class $i$ that are correctly predicted. The metrics are defined as follows:

- **Weighted Accuracy (WA)**: Calculated as the ratio of correctly predicted samples to the total number of samples:

$$WA = \frac{\sum_{i=1}^{C} n_{ii}}{N} \tag{12}$$

- **Unweighted Average Recall (UAR)**: The arithmetic mean of recall values for each class, providing an unbiased evaluation for imbalanced datasets:

$$UAR = \frac{1}{C} \sum_{i=1}^{C} \frac{n_{ii}}{N_i} \tag{13}$$

## The Prompts for SFT

"Step by step, identify the emotion expressed by the person in the video and support your answer with visual or auditory cues."

"Please think step by step and infer what emotion in the video represents."

"Please analyze the person's facial expression, tone of voice, and body language in the video to infer their emotional state. Explain your reasoning."

"Describe the mood of the individual in the video. What nonverbal signals or situational factors suggest this emotion?"

"Watch the video closely. What emotion is being portrayed, and how do the person' s actions, expressions, and environment contribute to your conclusion?"

Figure 9: Prompts of LLMs for logical-aware verification.

- **Weighted Average Recall (WAR)**: The average recall weighted by the number of samples in each class:

$$WAR = \sum_{i=1}^{C} \left( \frac{N_i}{N} \times \frac{n_{ii}}{N_i} \right) = \frac{\sum_{i=1}^{C} n_{ii}}{N} \tag{14}$$

# D Qualitative Analysis

To further demonstrate the effectiveness of **EmoOmni** in mitigating modality collapse and achieving precise emotion reasoning, we provide two additional qualitative examples in Figure 10 and 11.

These cases illustrate that the guided tokens in **OmniCoT** act as explicit cognitive anchors, forcing the model to "look" and "listen" before concluding, while **DyME-GRPO** adaptively balances these signals when certain modalities provide conflicting or ambiguous information.

# E Case Studies on CoT Paradigms

In Section 5.4, we discussed the necessity of Guided Tokens (GTs) by comparing OmniCoT with other common Chain-of-Thought (CoT) formats. To provide a more intuitive understanding of these structural differences, we present representative examples of different CoT paradigms in Figure 12. Existing benchmarks rely on unstructured, free-form reasoning, they lack explicit constraints for systematic evidence extraction across modalities. In contrast, OmniCoT utilizes Guided Tokens (GTs) as cognitive anchors, forcing the model to explicitly traverse specific modalities like `[vision]` and `[audio]` before reaching a conclusion, thereby ensuring transparency and modality-aware grounding.

# F Detailed Error Analysis on IEMOCAP

To understand the observed UAR fluctuations, we performed a per-class breakdown on the IEMOCAP dataset. Table 3 presents a detailed comparison between the SFT baseline and EmoOmni.

| Emotion | Count | OmniCoT-SFT | | EmoOmni (Ours) | |
|---|---|---|---|---|---|
| | | Prec. | Rec. | Prec. | Rec. |
| Neutral | 1555 | 76.78 | 69.74 | 52.22 | 71.83 |
| Happy | 560 | 75.36 | 78.43 | 44.77 | 88.75 |
| Sad | 999 | 27.53 | 28.38 | 88.80 | 22.22 |
| Anger | 1015 | 55.17 | 57.26 | 85.29 | 62.27 |
| Fear | 38 | 7.89 | 9.68 | 0.00 | 0.00 |
| Surprised | 98 | 9.19 | 24.31 | 48.00 | 12.24 |

Table 3: Per-class precision and recall (%) on IEMOCAP.

The per-class results reveal that EmoOmni's precision gains are concentrated in minority classes (*Sad*, *Anger*, and *Surprised*), suggesting more refined cross-modal discrimination. Although there is a slight precision trade-off in *Neutral* and *Happy*, the improved Recall indicates a transition from frequency-based heuristics to evidence-based reasoning.

The UAR fluctuation is primarily driven by the

**Ground_truth**: **Surprise**

**Qwen2.5-Omni-7b**:

Well, let's see from what I can tell about this scene with the woman looking at something off - screen while there is music playing.The setting seems to be indoors or maybe an office space given her attire which looks like business casual.Then without more context it s hard for me to pinpoint exactly one single feeling but if we have to choose based on common scenarios where such scenes might appear then perhaps she could feel surprised. Maybe because someone unexpected has just entered through that door behind her. So my answer would probably lean towards \"**surprise**\" .

**R1-Omni**:

<think>

In the video, a woman in a white shirt stands in front of a wall decorated with posters. Her facial expression is complex and subtle, with furrowed brows, slightly open eyes, downturned mouth, and slightly reddened pupils, revealing an inner confusion and unease. This emotion might be triggered by external stimuli, such as encountering unknown situations or receiving unexpected information. The entire scene conveys a sense of mystery and tension, prompting the audience to speculate about what has led to this particular emotional state.

</think>

<answer>**neutral**</answer>

**EmoOmni**:

<think>

[vision] The person in the image appears to be in a state of surprise or shock, with wide eyes and a slightly open mouth. There is no clear indication of happiness, sadness, anger, fear, disgust, or neutrality.

[audio] the sound seems to be a mix of surprise and fear.

[dialogue] None.

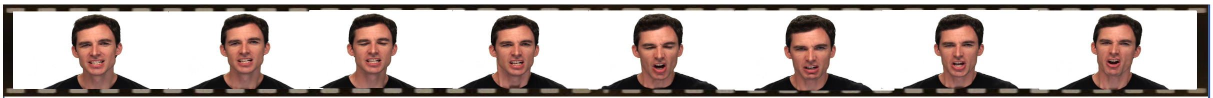
[analysis] the person's facial expression shows wide eyes and a slightly open mouth, which are typical signs of surprise or shock. additionally, the audio characteristics suggest a mix of surprise and fear, aligning with the visual cues. there is no indication of happiness, sadness, anger, fear, disgust, or neutrality based on the provided information.

</think>

<answer>**surprise**</answer>

Figure 10: Qualitative example from the MER2025 dataset.

*Fear* class (0.89% of the data). Unlike SFT, which may rely on stochastic noise for non-zero recall, DyME-GRPO prioritizes logical consistency over random guessing. Crucially, excluding this outlier raises the UAR to 51.46%, surpassing the SFT baseline. This proves that while our model excels in modality utilization, it remains sensitive to extreme long-tail distributions. Addressing such data scarcity is a key direction for future research.



**Ground_truth**: **Angry**

**Qwen2.5-Omni-7b**:

The man in the video is speaking in a calm and composed manner, with no signs of heightened emotions like rapid breathing or facial contortions that would suggest anger or fear. His tone is steady and measured, which is typical of a neutral emotion. Also, the content of the speech doesn't seem to evoke strong positive or negative reactions that would indicate happiness or sadness. It's just a straightforward statement without any emotional intensity. Answer: **Neutral**

**R1-Omni**:

<think>
The video features a man dressed in black, standing against a pure white background. His expression shifts from calm to serious and then to surprised, with his eyes and mouth moving as if speaking. This series of facial changes suggests he is experiencing an emotional fluctuation. The audio content provided by the video mentions \"Kids are talking by the door,\" indicating that the characters are engaged in a lively conversation. Combining these elements, it can be inferred that the man is facing an unexpected situation or having had an interesting exchange, triggering his heightened emotions.
</think>
<answer>**neutral**</answer>

**EmoOmni**:

<think>
[vision] The person in the image appears to be in a state of surprise or shock, with wide eyes and an open mouth. There is no clear indication of happiness, sadness, anger, fear, disgust, or neutrality.
[audio] the speaker's tone is harsh and forceful, which could indicate anger or fear.
[dialogue] Kids are talking by the door.
[analysis] the person's wide eyes and open mouth suggest surprise or shock rather than happiness, sadness, anger, fear, disgust, or neutrality. the harsh and forceful tone of the speaker might lean towards fear or anger, but the visual cues are more aligned with surprise.
</think>
<answer>**angry**</answer>

Figure 11: Qualitative example from the RAVDESS dataset.

| MERR |
| --- |
| <think><br>Lexical subtitle: Dogs are sitting by the door.<br>Audio tone description: speaks in a fearful tone.<br>Visual expression description: Lower eyebrowsandLips apart showing teeth<br>Visual objective description: The person in the image is a young man with a serious expression on his face. He is wearing a black shirt and has a black background behind him. The man is standing in front of a white wall, and his body language suggests that he is concentrating on something.<br>Multimodal description: In the video, a man's facial expressions, including lowered eyebrows and lips apart showing teeth, suggest a strong emotional response. His fearful tone and words, \"Dogs are sitting by the door,\" imply a sense of alarm or anxiety, indicating that he is likely feeling fearful.</think><br><answer>fearful<answer> |
| **MER 2025** |
| <think><br>In the video, a woman wearing traditional Chinese clothing is sitting on a chair with a man standing next to her. The woman's facial expression seems to be a mixture of surprise and joy, which can be observed from her eyes and the corners of her mouth. Her eyebrows are slightly raised and her lips are slightly open, indicating her reaction to something said or happening with the man next to her. On the other hand, the man is smiling, indicating that he may have shared a lighthearted and enjoyable moment with the woman. The surroundings are not very clear, but the presence of traditional Chinese clothing and the chair suggests that this scene may be a cultural event or formal gathering. The woman's reaction and the man's smile may be related to a joke or humorous situation, prompting the woman to display a mixture of surprise and joy. Overall, the woman and the man are enjoying a lighthearted and pleasant moment, sharing laughter and creating a positive atmosphere around them.In the audio, as an expert in the field of emotions, I can analyze the acoustic information in the audio to identify clues related to individual emotions. In this case, the individual's emotional state can be inferred through the acoustic information.<br></think><br><answer>surprise </answer> |
| **OmniCoT w/o GTs** |
| <think><br>The woman appears to be in a state of surprise or shock, as indicated by her wide eyes and slightly open mouth. There is no clear expression of happiness, sadness, anger, fear, disgust, or neutrality. the speaker's tone sounds apologetic and slightly distressed. Okay, I'm sorry. tthe woman's gentle touch and the babies' state of surprise or curiosity suggest a moment of unexpected interaction rather than fear or distress. the hint of surprise in the speaker's voice aligns with the babies' reaction, indicating a moment of surprise rather than sadness or anger<br></think><br><answer>neutral </answer> |

Figure 12: Qualitative comparison of different Chain-of-Thought (CoT) paradigms. The examples from **MERR** and **MER 2025** illustrate unstructured, free-form reasoning which often lacks systematic modality-specific grounding. The **OmniCoT w/o GTs** example demonstrates that even with high-quality reasoning text, the absence of explicit structural anchors (Guided Tokens) can lead to disorganized evidence integration and potential modality collapse, failing to resolve contradictions between visual and acoustic signals effectively.