EMO-R3: Reflective Reinforcement Learning for Emotional Reasoning in Multimodal Large Language Models

Yiyang Fang[1,2], Wenke Huang[1], Pei Fu[2*], Yihao Yang[1], Kehua Su[1], Zhenbo Luo[2], Jian Luan[2], Mang Ye[1†]
[1] School of Computer Science, Wuhan University.
[2] MiLM Plus, Xiaomi Inc.
{fangyiyang, yemang}@whu.edu.cn

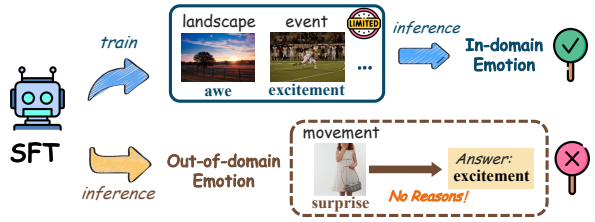
https://github.com/SeerRay-Lab/emo-r3

Abstract

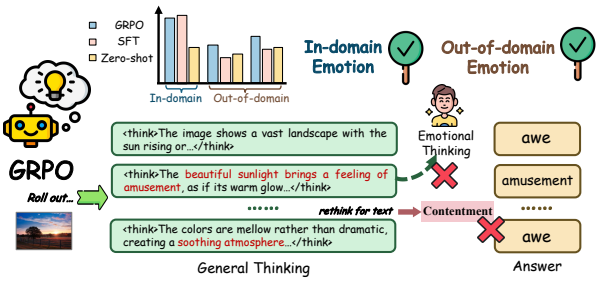
*Multimodal Large Language Models (MLLMs) have shown remarkable progress in visual reasoning and understanding tasks but still struggle to capture the complexity and subjectivity of human emotions. Existing approaches based on supervised fine-tuning often suffer from limited generalization and poor interpretability, while reinforcement learning methods such as Group Relative Policy Optimization fail to align with the intrinsic characteristics of emotional cognition. To address these challenges, we propose Reflective Reinforcement Learning for Emotional Reasoning (**EMO-R3**), a framework designed to enhance the emotional reasoning ability of MLLMs. Specifically, we introduce Structured Emotional Thinking to guide the model to perform step-by-step emotional reasoning in a structured and interpretable manner, and design a Reflective Emotional Reward that enables the model to re-evaluate its reasoning based on visual-text consistency and emotional coherence. Extensive experiments demonstrate that **EMO-R3** significantly improves both the interpretability and emotional intelligence of MLLMs, achieving superior performance across multiple visual emotional understanding benchmarks.*

1. Introduction

Multimodal Large Language Models (MLLMs) [22, 30] have achieved remarkable progress in visual question answering, visual understanding, and visual generation tasks by leveraging large-scale multimodal data [5, 24, 43, 51, 61]. However, despite their strong performance on general visual tasks, MLLMs still struggle to capture and interpret emotions effectively [56, 58], often generating superficial emotional responses and failing to fully understand complex emotional cues [6, 7, 53, 54, 65, 67].



(a) Limitation of supervised fine-tuning



(b) Non-generality of GRPO in emotional tasks

Figure 1. **Illustration of the motivation.** **(a)** SFT relies on human annotations but is constrained by fixed labels and limited categories, resulting in poor generalization and interpretability. It performs well on in-domain pairs like "landscape–awe" but struggles with out-of-domain or unseen cases (e.g., "movement-surprise"). **(b)** Although GRPO improves generalization, its think process is not emotion-oriented and weakly connected to the final answer (e.g., rethinking the last rollout yields "amusement", while the prediction is "fear").

In the field of visual emotional understanding, many existing studies such as EmoVIT [53], EmotionL-LaMA [6], AffectGPT [27], and EmoLLM [58] primarily adopt Supervised Fine-Tuning (SFT) [15, 16, 28] to improve model performance on emotional tasks. However, these approaches still have notable limitations in **generalization and interpretability** [40]. As shown in Fig. 1(a), SFT learns emo-

*Project Leader. Work done during internship at Xiaomi Inc.
†Corresponding Author.

tional representations by fitting the distribution of the training data, but the limited range of emotional categories and the fixed, predefined label taxonomy constrain the model to discrete emotional types. As a result, the model struggles to capture the continuity, subtle nuances, and contextual variability of visual emotional expressions. This reliance on a closed label space often leads to overfitting and reduces the adaptability of model to unseen visual or affective domains. Moreover, because SFT relies on example-level supervision, its reasoning tends to be pattern-matching rather than genuinely capturing the relationships among emotional factors. In contrast, applying Reinforcement Learning (RL) [31, 70] for post-training MLLMs can effectively alleviate these issues. In particular, Group Relative Policy Optimization (GRPO) [12, 41, 44, 46] stands out because, unlike Proximal Policy Optimization (PPO) [45] and Direct Preference Optimization (DPO) [39, 55], it does not require additional human-annotated reasoning traces for training [48]. Specifically, GRPO optimizes model behavior based on relative feedback among grouped samples, enabling the model to learn more generalizable emotional reasoning strategies through comparative evaluation. This optimization mechanism allows the model to uncover the latent structures and semantic relationships between visual content and emotional expressions, thereby enhancing its capability in visual emotional understanding and reasoning.

GRPO-based methods typically focus on optimizing the group-relative advantage [18, 63] or improving sampled roll-outs [4, 59, 60, 64] to enhance general capabilities, yet they pay little attention to task-specific adaptation for downstream emotional-understanding tasks. While recent emotion-related reinforcement learning works introduce GRPO into emotional reasoning [27, 69], they largely do so in a superficial manner, reusing its framework without adapting it to the intrinsic nature of emotional cognition. Actually, ❶ ***general GRPO generated reasoning process does not align well with the reasoning patterns required for emotion interpretation***, especially in visual emotion-understanding scenarios. While the decision-making of GRPO is effective, it fails to reliably capture the intuitive logic underlying human emotional comprehension.

Furthermore, unlike tasks such as mathematical reasoning [46] or code generation [42], where the relationship between *think* and *answer* is tightly bound, ❷ ***visual emotional understanding tasks lack this direct correspondence between reasoning traces and outputs***. In mathematical or programming tasks, an incorrect reasoning step almost inevitably leads to an incorrect answer, allowing GRPO to indirectly constrain the reasoning process through answer verification. In contrast, emotional understanding is highly subjective and context-dependent. The reasoning path may diverge from the final *answer* due to individual or contextual variations in emotional interpretation. As shown in Fig. 1(b), when we *rethink* the *think*-text from roll-out samples, the inferred emotion often differs from that of the final *answer*, indicating that the correctness of the answer cannot reliably reflect the quality of the reasoning process. Visual emotional tasks require not only perception of visual cues but also comprehension of complex emotional contexts and background knowledge, while maintaining emotional coherence across these cues. Therefore, constraining the answer alone is insufficient to guide the reasoning process effectively, posing a unique challenge for enhancing emotional reasoning in vision-based tasks.

To tackle these challenges, we propose **R**eflective **R**einforcement Learning for Emotional **R**easoning in Multimodal Large Language Models (`EMO-R3`). ***First***, we design Structured Emotional Thinking that explicitly guides the model to reason about emotions in a step-by-step manner and constrains its output to follow a specific, interpretable format. This structured formulation helps the model generate coherent emotional reasoning traces rather than fragmented or task-agnostic thoughts. ***Next***, we introduce Reflective Emotional Reward, which allows the model to re-evaluate its own reasoning and assess whether its emotional interpretation aligns with visual and contextual cues. By feeding the reasoning back into the model, we apply two rewards: visual-text consistency, ensuring the reasoning is grounded in the visual input, and emotional reasoning validity, enforcing logical soundness and emotional coherence in the inferred emotions. Extensive experiments demonstrate that `EMO-R3` significantly enhances the interpretability and emotional intelligence of multimodal large language models in visual emotional understanding.

The main contributions can be summarized as follows:

- We propose a Structured Emotional Thinking process that guides MLLMs to perform emotional reasoning in a structured and interpretable manner, improving their ability to understand emotions more human-likely.
- We introduce a Reflective Emotional Reward mechanism that enables the model to re-evaluate its reasoning and optimize through reflective feedback, ensuring more coherent and grounded emotional reasoning.
- We conduct extensive experiments demonstrating that `EMO-R3` consistently outperforms previous methods from multiple perspectives.

2. Related Works

2.1. Emotion Recognition in MLLMs

Recent advances in Multimodal Large Language Models (MLLMs) [5, 22, 24, 30, 51] have significantly enhanced the joint understanding across visual, textual, and auditory modalities [20, 21, 33], and improved the ability to handle a variety of multimodal tasks [15, 16, 28]. Most research in this field focuses on leveraging large-scale pretrained mod-

els for general-purpose applications [2, 13, 32, 68], including vision-language reasoning [35, 36, 50], image captioning [29, 62], and visual question answering [11, 17, 19, 47]. MLLMs have demonstrated remarkable performance in these tasks [3], showcasing their ability to integrate and reason across multiple modalities.

However, MLLMs often struggle with emotion-related tasks [27]. These challenges arise from the subjective and context-dependent nature of emotional understanding, which requires not only perceptual grounding but also affective reasoning across modalities. To address this issue, several studies have explored supervised fine-tuning MLLMs using emotional datasets [54, 58, 65, 67]. For example, EmoVIT [53] leverages GPT-4 to generate emotion-relevant textual descriptions, helping models better interpret affective cues and capture nuanced emotional expressions. Meanwhile, Emotion-LLaMA [6] integrates specialized affective encoders that are designed to capture and interpret emotional signals across multiple modalities, thereby enhancing the capacity of model to understanding emotions from text, audio, and visual inputs. Although these fine-tuning approaches improve performance, they typically require extensive retraining or instruction-based adaptation, leading to high computational costs and limited scalability. Nowadays, increasing attention has been devoted to enhancing the generalization and interpretability of MLLMs [25, 26, 52], which has motivated the exploration of reinforcement learning strategies for downstream emotional understanding. These approaches aim to improve the affective cognition of models and human alignment in open-domain scenarios through more flexible reward signals and adaptive optimization processes.

2.2. Group Relative Policy Optimization

With the growing adoption of reinforcement learning [31, 70] in large language models (LLMs) [49] training [10, 66], Group Relative Policy Optimization (GRPO) [41, 46] has emerged as a widely used optimization paradigm, originally applied to enhance reasoning and alignment performance in LLMs. Unlike traditional Proximal Policy Optimization (PPO) [45], GRPO optimizes based on in-group relative rewards, generating multiple candidate reasoning trajectories for the same input and computing relative advantages among them. This formulation greatly improves optimization stability and reasoning consistency. Early works such as the DeepSeek-R1 [12] demonstrated that GRPO can substantially enhance model performance and interpretability in mathematical, logical, and scientific reasoning tasks.

Recently, extensive studies have extended and refined the GRPO framework. For instance, Text-Debiased Hint-GRPO [14] introduces debiased hint mechanisms to mitigate linguistic bias in multimodal reasoning; R1-VL [64] adopts a step-wise optimization strategy to stabilize learning across multimodal tasks; and R1-Omni [69] applies reinforcement learning to omni-modal emotion recognition, verifying its potential in subjective affective reasoning. Moreover, Video-R1 [9], VideoChat-R1 [23], and Visual-RFT [34] extend the paradigm to video and vision-centric settings, showcasing its cross-modal scalability.

Nevertheless, in the field of emotion understanding, the application of GRPO remains largely superficial. Most methods merely adapt the general GRPO framework at a surface level [27, 69], without addressing its inherent mismatch with subjective emotional reasoning. Specifically, the reasoning paths generated by general GRPO often diverge from the intuitive logic of human affective reasoning, making it difficult to capture subjective and context-dependent emotional associations. Furthermore, unlike mathematical or coding tasks, where the thinking–answer relationship is tightly coupled, visual emotion understanding lacks such direct correspondence, causing traditional GRPO to struggle with learning stable affective semantic signals. Therefore, developing a GRPO optimization mechanism tailored to emotional understanding is essential for advancing affective reasoning in multimodal large models.

3. The Proposed Method

3.1. Preliminary

Group Relative Policy Optimization (GRPO) is a variant of Proximal Policy Optimization (PPO). Originally, PPO was designed to enhance mathematical reasoning in large language models. However, GRPO can be effectively adapted to improve visual reasoning and other multimodal capabilities as well. GRPO begins by constructing the current policy model π_θ and a reference model π_{old}, where the latter represents the *old* policy, i.e., the policy from a previous iteration. Let ρ_Q denote the distribution of prompts or questions. Given a prompt $q \sim \rho_Q$, the model samples a group of outputs $o_1, o_2, \ldots, o_G$ from the old policy π_{old}. The policy π_θ is then optimized by maximizing the following objective function:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim\rho_Q}\mathbb{E}_{o\sim\pi_{\text{old}}(\cdot|q)}\left[\frac{1}{G}\sum_{i=1}^{G} f_\epsilon\left(\frac{\pi_\theta(o_i|q)}{\pi_{\text{old}}(o_i|q)}, \hat{A}_i\right)\right] - \beta\,\mathbb{D}_{KL}[\pi_\theta\|\pi_{\text{ref}}]\,, \tag{1}$$

where β is the hyperparameter, and $f_\epsilon(x, y) = \min(xy, \text{clip}(x, 1-\epsilon, 1+\epsilon)y)$. $\hat{A}i$ denotes the advantage, which is calculated based on the relative rewards of the outputs within each group. More specifically, for each question q, a group of outputs $o_1, o_2, \ldots, o_G$ is sampled from the old policy model π_{old}. A reward function R is then used to score these outputs, yielding G rewards $\mathbf{r} = r_1, r_2, \ldots, r_G$, where $r_i = \mathcal{R}(q, o_i)$. The mean reward is computed as

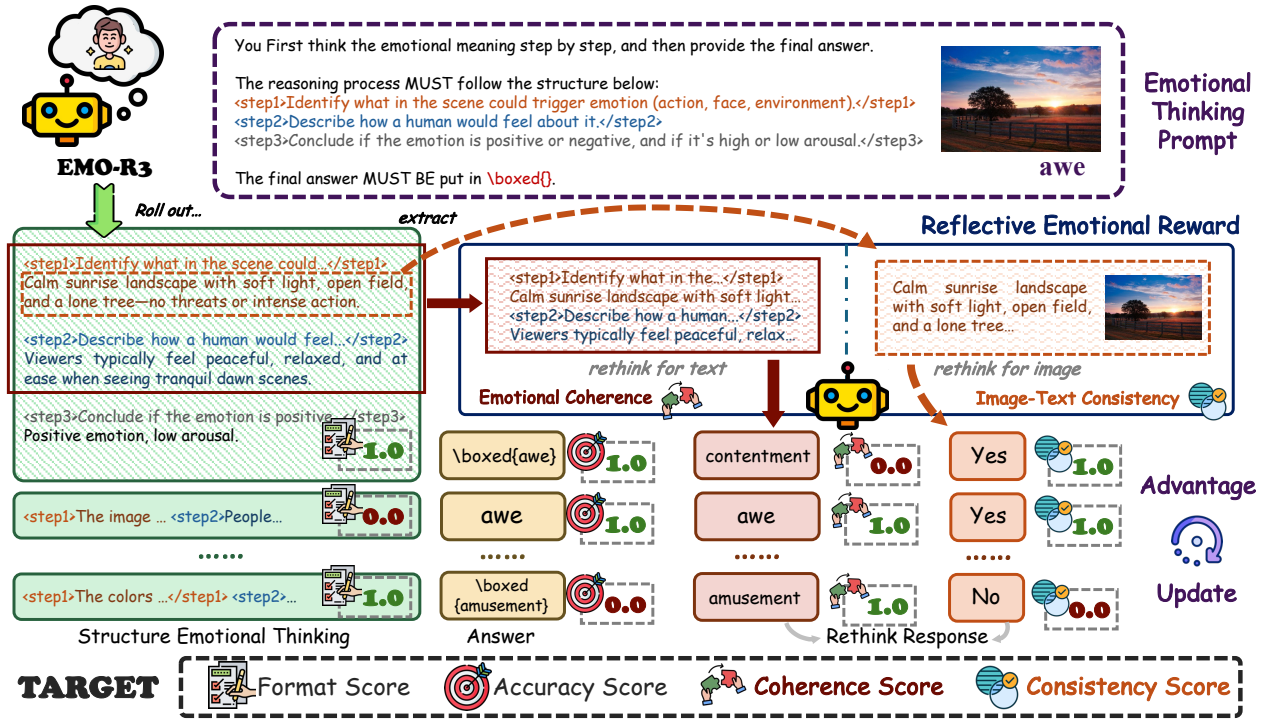


Figure 2. **Architecture illustration** of EMO-R3. The upper part presents the Structured Emotional Thinking prompt, which consists of three consecutive thinking steps followed by a final answer. The lower part illustrates the Reflective Emotional Reward mechanism, where multiple rollout samples are evaluated based on image–text consistency and emotional coherence, and are jointly optimized with the original Format and Accuracy rewards under the GRPO framework.

$\mu = \frac{1}{G}\sum_{i=1}^{G} r_i$, and the standard deviation is defined as $\sigma = \sqrt{\frac{1}{G}\sum_{i=1}^{G}(r_i - \mu)^2}$. The normalized advantage for the i^{th} rollout is then defined as $\hat{A}_i = \frac{r_i - \mu}{\sigma}$. This normalization ensures that the advantage values have zero mean and unit variance within each group, stabilizing gradients and promoting consistent optimization dynamics.

3.2. Structured Emotional Thinking (SET)

Motivation. Although GRPO has shown effectiveness in improving general reasoning abilities of large multimodal models, its prompting design for the thinking stage is often minimal, typically consisting of a single instruction such as *think*. This one-step thinking cue is task-agnostic and provides no explicit guidance on how emotional reasoning should be organized. In emotional understanding tasks, such a simplistic prompt frequently causes the model to generate fragmented or inconsistent emotional reasoning traces that fail to capture the subtle relationships between visual cues and human affective appraisal. In visual emotional understanding, where the mapping between perception and emotion is complex and context dependent, a single *think* instruction is insufficient to elicit coherent or human-aligned reasoning.

Designed Prompt. To achieve interpretable and human-like emotional understanding, we propose Structured Emotional Thinking (SET), a module that guides the model to perform emotion reasoning in a structured, step-by-step manner before generating the final prediction.

Concretely, SET constrains the reasoning process into three explicit stages, mirroring how humans interpret emotions in visual scenes:

Structured Emotional Thinking:

- **Emotional Trigger Identification:** Detect which elements in the scene (objects, actions, environments, or facial cues) may trigger emotional responses.
- **Human Emotional Reflection:** Describe how a human observer would emotionally respond to these elements.
- **Emotional Conclusion:** Determine whether the overall emotion is positive or negative, and assess its arousal level (e.g., calm vs. excited).

Given a multimodal input pair (I, T), the model generates a structured reasoning output $o = \{s_1, s_2, s_3, \hat{\mathcal{E}}\}$ corresponding to the three stages and the final *answer* of emotional reasoning:

$$o = \mathcal{M}_\theta(I, T), \quad \hat{\mathcal{E}} = \mathcal{F}_a(o), \tag{2}$$

where $\mathcal{M}_\theta$ denotes the multimodal reasoning model parameterized by θ, and $\mathcal{F}_a(\cdot)$ outputs the final *answer* enclosed in `\boxed{}`.

General Reward. Following the GRPO setting, we define two reward terms to guide the optimization of the structured emotional reasoning model.

The format reward $\mathcal{R}_{\text{format}}$ measures whether the generated reasoning sequence adheres to the prescribed structure. Specifically, it checks whether each reasoning step s_i corresponds to the expected stage and whether the final *answer* is correctly enclosed in `\boxed{}`:

$$\mathcal{R}_{\text{format}} = \begin{cases} 1, & \text{if the } \texttt{step} \text{ and } \texttt{box} \text{ format are correct;} \\ 0, & \text{otherwise.} \end{cases}$$

Meanwhile, the accuracy reward $\mathcal{R}_{\text{acc}}$ evaluates whether the predicted emotional label $\hat{\mathcal{E}}$ aligns with the ground-truth emotion label $\mathcal{E}^*$:

$$\mathcal{R}_{\text{acc}} = \begin{cases} 1, & \text{if } \hat{\mathcal{E}} = \mathcal{E}^*; \\ 0, & \text{otherwise.} \end{cases}$$

These two general rewards serve as the foundational supervision signal that initiates the optimization of the structured emotional thinking model under the GRPO framework, ensuring that the model first learns to produce structurally valid and semantically accurate emotional reasoning before incorporating higher-level reflective objectives.

3.3. Reflective Emotional Reward (RER)

Motivation. Although the Designed Prompt provides a structured framework for emotional reasoning, it cannot ensure that the generated reasoning is visually consistent with the textual content or emotionally coherent. Meanwhile, the general GRPO formulation lacks effective constraints on the *think* process; by supervising only the final *answer*, it fails to effectively select high-quality reasoning samples.

Image-Text Consistency Reward. The image-text consistency reward enforces alignment between the generated reasoning and the visual content of the image.

In this process, we extract only `step1` from the model output o, denoted as $s_1 = \mathcal{F}_1(o)$, and feed it back into the model together with the image I. The prompt for this reflective process is denoted as $\mathcal{P}_{\text{cons}}$:

> ***Can the following text describe the image?***

The model then produces a reflective output:

$$\hat{y}_{\text{cons}} = \mathcal{M}(I, s_1, \mathcal{P}_{\text{cons}}), \tag{3}$$

Algorithm 1: `EMO-R3`

Input: Dataset $\mathcal{D} = \{(I, T, \mathcal{E}^*)\}$, pretrained model $\mathcal{M}_\theta$, rollout number G, coefficients λ_1, λ_2
Output: Optimized model $\mathcal{M}'_\theta$
foreach $(I, T, \mathcal{E}^*) \in \mathcal{D}$ **do**
  /* Generate multiple reasoning outputs with structured prompt */
  $\{o_1, o_2, \ldots, o_G\} \sim \pi_{\text{old}}(\cdot|I, T)$
  /* Compute rewards for each rollout */
  for $i = 1$ **to** G **do**
    /* (1) General Reward */
    Parse $o_i = \{s_1, s_2, s_3, \hat{\mathcal{E}}\}$
    $\mathcal{R}^{(i)}_{\text{format}} = \mathbb{I}(\text{step and box format correct})$
    $\mathcal{R}^{(i)}_{\text{acc}} = \mathbb{I}(\hat{\mathcal{E}} = \mathcal{E}^*)$
    /* (2) Reflective Emotional Reward */
    Image-text consistency:
    $\hat{y}^{(i)}_{\text{cons}} = \mathcal{M}(I, s_1, \mathcal{P}_{\text{cons}})$, $\mathcal{R}^{(i)}_{\text{cons}} = \mathbb{I}(\hat{y}^{(i)}_{\text{cons}} = \text{Yes})$
    Emotional coherence:
    $\hat{y}^{(i)}_{\text{coh}} = \mathcal{M}(s_{1,2}, \mathcal{P}_{\text{coh}})$, $\mathcal{R}^{(i)}_{\text{coh}} = \mathbb{I}(\hat{y}^{(i)}_{\text{coh}} = \mathcal{E}^*)$
    $\mathcal{R}^{(i)}_{\text{RER}} = \frac{1}{2}(\mathcal{R}^{(i)}_{\text{cons}} + \mathcal{R}^{(i)}_{\text{coh}})$
    /* (3) Combine into overall reward */
    $\mathcal{R}^{(i)}_{\text{overall}} = (1-\lambda_1-\lambda_2)\mathcal{R}^{(i)}_{\text{acc}} + \lambda_1 \mathcal{R}^{(i)}_{\text{RER}} + \lambda_2 \mathcal{R}^{(i)}_{\text{format}}$
  end
  /* Compute advantage and update model via GRPO */
  Normalize rewards within group to obtain $\hat{A}_i$;
  Update policy parameters θ using GRPO objective with advantages $\hat{A}_i$.
end
return $\mathcal{M}'_\theta$

where the response can be either "Yes" or "No". The corresponding reward is defined as:

$$\mathcal{R}_{\text{cons}} = \begin{cases} 1, & \text{if } \hat{y}_{\text{cons}} = \text{Yes}; \\ 0, & \text{if } \hat{y}_{\text{cons}} = \text{No}. \end{cases}$$

This reward encourages the model to generate reasoning that is both emotionally coherent and visually grounded, ensuring stronger alignment between textual descriptions and image semantics.

Emotional Coherence Reward. The emotional coherence reward aims to evaluate whether the reasoning process maintains consistency with the ground-truth emotion label.

In this process, we extract `step1` and `step2` from the model-generated reasoning, denoted as $s_{1,2} = \mathcal{F}_{1,2}(o)$, which is fed back into the model for reflection. The prompt for this reflective process is denoted as $\mathcal{P}_{\text{coh}}$:

> ***Which emotion best describes the text above?***

The model then produces a reflective output:

$$\hat{y}_{\text{coh}} = \mathcal{M}(R_{\text{input}}, \mathcal{P}_{\text{coh}}), \tag{4}$$

where $\hat{y}_{\text{coh}}$ represents the emotion label predicted by the model during the reflective stage. We then compare $\hat{y}_{\text{coh}}$ with the ground-truth emotion label $\mathcal{E}^*$. The emotional coherence reward is defined as:

$$\mathcal{R}_{\text{coh}} = \begin{cases} 1, & \text{if } \hat{y}_{\text{coh}} = \mathcal{E}^*; \\ 0, & \text{otherwise.} \end{cases}$$

This reward encourages the model to generate reasoning that is emotionally consistent with the ground-truth label, thereby improving the emotional coherence and interpretability of the reasoning process.

3.4. Overall Reward and Discussion

Overall Reward. The final optimization objective integrates all the previously defined reward components into a unified formulation. Specifically, the reflective emotional reward is obtained by averaging the image-text consistency reward and the emotional coherence reward:

$$\mathcal{R}_{\text{RER}} = \frac{\mathcal{R}_{\text{cons}} + \mathcal{R}_{\text{coh}}}{2}. \tag{5}$$

Subsequently, the overall reward used for GRPO optimization is defined as a weighted combination of the accuracy reward, the reflective emotional reward, and the format reward, which is calculated as follows:

$$\mathcal{R}_{\text{overall}} = (1 - \lambda_1 - \lambda_2)\, \mathcal{R}_{\text{acc}} + \lambda_1\, \mathcal{R}_{\text{RER}} + \lambda_2\, \mathcal{R}_{\text{format}}, \tag{6}$$

where λ_1 and λ_2 are balancing coefficients that control the relative contributions of emotional coherence and structural correctness.

By combining these complementary rewards, the training process promotes reasoning traces that are not only visually grounded and emotionally coherent but also maintain interpretable, human-aligned structure.

Discussion on Cold-Start-Emo. In our framework, a key question is whether Supervised Fine-Tuning (SFT) should be introduced as a cold-start stage before GRPO optimization. Unlike factual reasoning or visual question answering tasks, emotion recognition inherently involves subjectivity. Pretrained MLLMs often carry emotional priors derived from large-scale corpora, which reflect general or culture-dependent affective tendencies. These priors may deviate considerably from the labeling schemes of specific downstream datasets. If GRPO training is conducted without any prior alignment, such a mismatch can cause the model to repeatedly generate reasoning traces that are inconsistent with the dataset annotations, leading to sparse reward signals and consequently weakening optimization stability.

Inspired by previous studies, many works perform cold start with Chain-of-Thought (CoT)-annotated datasets prior to GRPO. The main motivation behind this design is to endow the model with an initial *thinking* ability, enabling more effective reasoning optimization later. Our motivation, however, is different: instead of enhancing the reasoning-chain capability of model, we aim to alleviate the training difficulty caused by subjective bias in emotional understanding tasks.

To this end, we explore a lightweight SFT-based Cold Start for Emotional Reasoning (Cold-Start-Emo) using a small number of samples without CoT annotations. This stage requires no additional reasoning chains; rather, a small set of task-specific examples is used to help the model preliminarily learn the task format, emotional label system, and expression patterns. Such initialization enables an early-stage alignment between the pretrained priors and the target task distribution. Empirical results demonstrate that this initialization allows the model to generate higher-quality rollouts more stably during subsequent GRPO training, mitigating reward sparsity and ultimately improving both the coherence and accuracy of emotional reasoning.

4. Experiments

4.1. Experimental Setup

Environment and Dataset. Our training framework uses EasyR1, and the testing framework uses NoisyRollout. Our experiments use three emotion datasets: EmoSet [57] (8 categories), Emotion6 [38] (6 categories), and WebEmo [37] (7 categories). We train separately on the EmoSet and Emotion6 datasets, while other datasets are used as out-of-domain tests. To enhance efficiency, we randomly cropped the training and testing sets of each dataset (2,000 samples).

Architecture and Counterparts. We utilize the popular open-source Qwen2.5-VL-3B-Instruct [1] as the base (Vanilla) model, which exhibits strong foundational capabilities well-suited for subsequent RL training. We further compare our approach with the training-free method SEPM [8], as well as GRPO [46] and DAPO [63], to validate its effectiveness.

Implement Details. The experimental results are obtained at the same step (when convergence is reached), with the hyperparameters λ_1 and λ_2 both set to 0.1. The learning rate for the experiment is set to 2.0e-6. To eliminate randomness, we ran the experiment three times and reported the median result. The experiments are conducted on a total of 8 NVIDIA H20 GPUs, each with 96GB of memory.

Evaluation Metrics. We evaluate both in-domain and out-of-domain performance. For each dataset, we use Accuracy (ACC) as the evaluation metric. We further compute the average performance for in-domain ($\mathcal{A}^I$) and out-

Table 1. **Performance comparison with the state-of-the-art GRPO variants** on the emotional reasoning tasks across in-domain and out-of-domain settings. * denotes models without post-training. Datasets marked with the superscript I, *e.g.* EmoSet^I and Emotion6^I, denote the in-domain training dataset. We mark the Best in bold across different methods. Please refer to Sec. 4.2 for details.

Methods	Roll-out	EmoSet^I	Emotion6	WebEmo	Emotion6^I	EmoSet	WebEmo	$\mathcal{A}^I$	$\mathcal{A}^O$	$\mathcal{A}$
LLaVA1.5-7B										
Vanilla*	-	52.77	48.32	25.56	48.32	52.77	25.56	50.55	38.05	42.22
SEPM*	-	56.04	54.21	42.39	54.21	56.04	42.39	55.13	48.76	50.88
Qwen2.5-VL-3B-Instruct										
Vanilla*	-	51.55	50.00	40.65	50.00	51.55	40.65	50.77	45.71	47.40
SFT	-	77.15	34.51	17.75	69.53	26.45	37.65	73.34	29.09	43.84
GRPO		74.60	60.10	49.50	70.88	59.90	44.85	72.74	53.59	59.97
DAPO	*4*	68.99	56.90	49.80	68.56	59.95	45.50	68.78	53.04	58.28
EMO-R3		75.50	60.44	50.45	70.71	60.70	45.20	**73.10**	**54.20**	**60.50**
GRPO		75.45	57.91	49.40	69.87	60.30	42.05	72.66	52.42	59.16
DAPO	*8*	70.21	55.72	48.80	62.39	58.05	46.30	66.30	52.22	56.91
EMO-R3		76.40	59.26	49.70	71.72	61.80	43.65	**74.06**	**53.60**	**60.42**

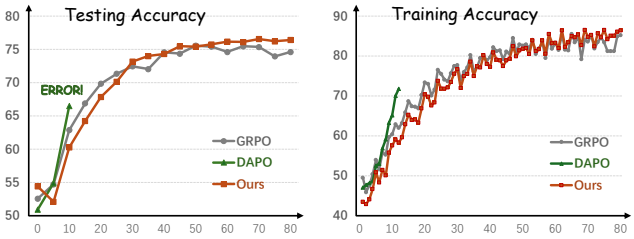


Figure 3. Training and testing accuracy during the training process. DAPO fails to conduct complete training. A more detailed analysis of this failure is provided in Sec. 4.2.

of-domain ($\mathcal{A}^O$) evaluations. Finally, we take the average of all these results to obtain the overall performance ($\mathcal{A}$).

4.2. Comparison Experiments

Comparison with State-of-the-art. We compare the proposed approach with GRPO variants on both in-domain and out-of-domain emotional datasets, as well as with several training-free methods. As reported in Tab. 1, EMO-R3 consistently achieves the highest overall accuracy across both 4-rollout and rollout-8 settings, demonstrating its ability to enhance emotional reasoning. Compared to these baselines, our method yields higher in-domain performance, reflecting better alignment with emotional cues learned from the training distributions. Meanwhile, the gain in out-of-domain accuracy shows that our learning strategy mitigates overfitting and improves robustness to domain shift. Notably, DAPO fails to conduct complete training, as shown in Fig. 3. The failure stems from a fundamental mismatch between the filtering strategy of DAPO and the discrete nature of emotional reasoning evaluation, where the binary reward structure conflicts with the continuous filtering criteria, leading to sample depletion and training instability.

Experiment on Cold-Start-Emo. We explore the Cold-Start-Emo under the rollout-8 setting. The Cold-Start-Emo is designed to provide early alignment and stabilize the learning process for emotional reasoning. As shown in Tab. 2, the integration of Cold-Start-Emo significantly outperforms EMO-R3 and all other baselines on the in-domain dataset, and it achieves the highest overall average accuracy on the out-of-domain datasets. This empirical evidence validates that Cold-Start-Emo is a highly effective initialization strategy that generates higher-quality rollouts and mitigating reward sparsity during subsequent GRPO training.

Table 2. **Experiment on Cold-Start-Emo.** EMO-R3$^\#$ denotes EMO-R3 with additional Cold-Start-Emo module. See Sec. 4.2.

Methods	EmoSet^I	Emotion6	WebEmo	$\mathcal{A}$
SFT	77.15	34.51	17.75	43.14
GRPO	75.45	57.91	49.40	60.92
EMO-R3	76.40	**59.26**	49.70	61.79
EMO-R3$^\#$	**77.81**	58.59	**50.00**	**62.13**

Table 3. **Ablative study** of Structured Emotional Thinking (SET) and Reflective Emotional Reward (RER). Please see Sec. 4.3.

SET	RER	EmoSet^I	Emotion6	WebEmo	$\mathcal{A}$
○	○	75.45	57.91	49.40	60.92
●	○	76.02	58.42	49.20	61.21
●	●	76.40	59.26	49.70	61.79

4.3. Ablation Experiments

In Tab. 3, we begin by validating the effectiveness of our proposed components through their incremental integration. As shown, incorporating Structured Emotional Thinking (SET) consistently enhances performance compared with the baseline, indicating that explicitly organizing the emotional reasoning procedure helps the model produce more coherent, fine-grained, and interpretable emotion represen-

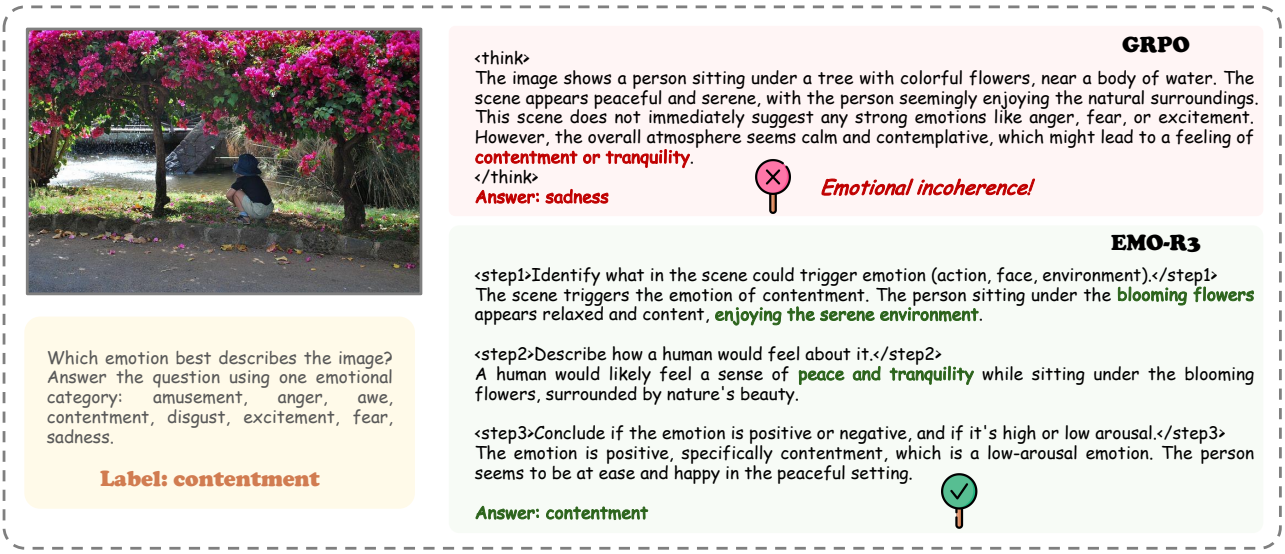


Figure 4. **Case study** between GRPO and EMO-R3 on the EmoSet dataset. Please see Sec. 4.4 for details.

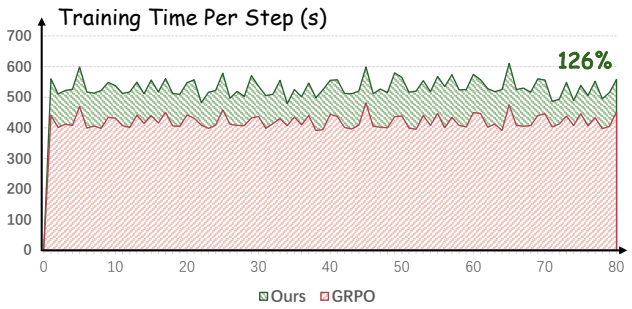


Figure 5. **Efficiency analysis** on the training process. See Sec. 4.5.

tations. When the Reflective Emotional Reward (RER) is further introduced, the model achieves additional improvements, suggesting that reflective self-assessment encourages the model to better align its emotional reasoning with the underlying multimodal evidence. Taken together, these findings demonstrate that the combined use of SET and RER not only improves the interpretability of the reasoning process but also substantially enhances the emotional intelligence of MLLMs.

4.4. Case Study

We evaluated the methods on the EmoSet dataset and selected a representative case for detailed analysis. We observed that the naive GRPO failed to attend to the most emotionally salient regions (*blooming flowers*), and its think and answer components exhibited emotional incoherence. In contrast, our proposed method (EMO-R3) effectively addresses this issue by producing emotionally coherent reasoning and predictions. This case demonstrates that EMO-R3 can accurately capture subtle affective cues and exhibit emotionally coherent reasoning, thereby leading to better emotional understanding and overall performance.

4.5. Efficiency Analysis

Considering that we introduce an additional reflection stage, we conducts an efficiency analysis on the training process under the rollout-8 setting on the EmoSet dataset. As shown in Fig. 5, although our method introduces a certain amount of extra computation time, it does not lead to a proportional increase in training cost. Moreover, our inference process does not require the reflection module, so it introduces **no additional inference-time cost**. Thus, our approach maintains high efficiency while achieving better performance.

5. Conclusion

In this work, we investigate the challenges of interpretability and generalization faced by Multimodal Large Language Models (MLLMs) in emotional understanding. Although general GRPO-based approaches can partially alleviate these issues, existing models still struggle to accurately capture the subtle, subjective, and context-dependent nature of human emotions. To address this gap, We propose Reflective Reinforcement Learning for Emotional Reasoning in Multimodal Large Language Models (EMO-R3). Our method integrates a Structured Emotional Thinking module to guide step-by-step affective reasoning and employs a Reflective Emotional Reward mechanism to ensure visual–textual consistency and coherent emotional expression. Without requiring additional annotations, EMO-R3 significantly improves both the interpretability and generalization of MLLMs. We believe this work offers new insights for developing emotionally intelligent and human-aligned MLLMs. Building upon this foundation, future research may further explore the generalization of emotion recognition with reasoning in more complex multimodal scenarios, including sequential or interactive task settings.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 6

[2] Yang Bai, Yucheng Ji, Min Cao, Jinqiao Wang, and Mang Ye. Chat-based person retrieval via dialogue-refined cross-modal alignment. In *CVPR*, 2025. 3

[3] Jinhe Bi, Yujun Wang, Haokun Chen, Xun Xiao, Artur Hecker, Volker Tresp, and Yunpu Ma. Visual instruction tuning with 500x fewer parameters through modality linear representation-steering. *arXiv preprint arXiv:2412.12359*, 2024. 3

[4] Minghan Chen, Guikun Chen, Wenguan Wang, and Yi Yang. Seed-grpo: Semantic entropy enhanced grpo for uncertainty-aware policy optimization. *arXiv preprint arXiv:2505.12346*, 2025. 2

[5] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, pages 24185–24198, 2024. 1, 2

[6] Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Jingdong Sun, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander Hauptmann. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. In *NeurIPS*, 2024. 1, 3

[7] Yiyang Fang, Wenke Huang, Guancheng Wan, Kehua Su, and Mang Ye. Emoe: Modality-specific enhanced dynamic emotion experts. In *CVPR*, 2025. 1

[8] Yiyang Fang, Jian Liang, Wenke Huang, He Li, Kehua Su, and Mang Ye. Catch your emotion: Sharpening emotion perception in multimodal large language models. In *ICML*, 2025. 6

[9] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*, 2025. 3

[10] Jiaxuan Gao, Shusheng Xu, Wenjie Ye, Weilin Liu, Chuyi He, Wei Fu, Zhiyu Mei, Guangju Wang, and Yi Wu. On designing effective rl reward at training time for llm reasoning. *arXiv preprint arXiv:2410.15115*, 2024. 3

[11] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *CVPR*, 2017. 3

[12] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 2, 3

[13] Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. Onellm: One framework to align all modalities with language. In *CVPR*, pages 26584–26595, 2024. 3

[14] Qihan Huang, Weilong Dai, Jinlong Liu, Wanggui He, Hao Jiang, Mingli Song, Jingyuan Chen, Chang Yao, and Jie Song. Boosting mllm reasoning with text-debiased hint-grpo. *arXiv preprint arXiv:2503.23905*, 2025. 3

[15] Wenke Huang, Jian Liang, Xianda Guo, Yiyang Fang, Guancheng Wan, Xuankun Rong, Chi Wen, Zekun Shi, Qingyun Li, Didi Zhu, et al. Keeping yourself is important in downstream tuning multimodal large language model. *arXiv preprint arXiv:2503.04543*, 2025. 1, 2

[16] Wenke Huang, Jian Liang, Zekun Shi, Didi Zhu, Guancheng Wan, He Li, Bo Du, Dacheng Tao, and Mang Ye. Learn from downstream and be yourself in multimodal large language model fine-tuning. In *ICML*, 2025. 1, 2

[17] Wenke Huang, Jian Liang, Guancheng Wan, Didi Zhu, He Li, Jiawei Shao, Mang Ye, Bo Du, and Dacheng Tao. Be confident: Uncovering overfitting in mllm multi-task tuning. In *ICML*, 2025. 3

[18] Wenke Huang, Quan Zhang, Yiyang Fang, Jian Liang, Xuankun Rong, Huanjin Yao, Guancheng Wan, Ke Liang, Wenwen He, Mingjun Li, et al. Mapo: Mixed advantage policy optimization. *arXiv preprint arXiv:2509.18849*, 2025. 2

[19] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, pages 6700–6709, 2019. 3

[20] Can Jin, Tong Che, Hongwu Peng, Yiyuan Li, Dimitris Metaxas, and Marco Pavone. Learning from teaching regularization: Generalizable correlations should be easy to imitate. *NeurIPS*, 37:966–994, 2024. 2

[21] Can Jin, Hongwu Peng, Qixin Zhang, Yujin Tang, Dimitris N Metaxas, and Tong Che. Two heads are better than one: Test-time scaling of multi-agent collaborative reasoning. *arXiv preprint arXiv:2504.09772*, 2025. 2

[22] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 1, 2

[23] Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*, 2025. 3

[24] Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. In *CVPR*, pages 26763–26773, 2024. 1, 2

[25] Zheng Lian, Haiyang Sun, Licai Sun, Hao Gu, Zhuofan Wen, Siyuan Zhang, Shun Chen, Mingyu Xu, Ke Xu, Kang Chen, et al. Explainable multimodal emotion recognition. *arXiv preprint arXiv:2306.15401*, 2023. 3

[26] Zheng Lian, Licai Sun, Mingyu Xu, Haiyang Sun, Ke Xu, Zhuofan Wen, Shun Chen, Bin Liu, and Jianhua Tao. Explainable multimodal emotion reasoning. *CoRR*, 2023. 3

[27] Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, Yong Ren, Zebang Cheng, Bin Liu, Rui Liu, Xiaojiang Peng, et al. Affectgpt: A new dataset, model, and benchmark for emotion understanding with multimodal large language models. *arXiv preprint arXiv:2501.16566*, 2025. 1, 2, 3

[28] Jian Liang, Wenke Huang, Guancheng Wan, Qu Yang, and Mang Ye. Lorasculpt: Sculpting lora for harmonizing general and specialized knowledge in multimodal large language models. In *CVPR*, 2025. 1, 2

[29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014. 3

[30] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2023. 1, 2

[31] Keliang Liu, Dingkang Yang, Ziyun Qian, Weijie Yin, Yuchi Wang, Hongsheng Li, Jun Liu, Peng Zhai, Yang Liu, and Lihua Zhang. Reinforcement learning meets large language models: A survey of advancements and applications across the llm lifecycle. *arXiv preprint arXiv:2509.16679*, 2025. 2, 3

[32] Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. *arXiv preprint arXiv:2402.09353*, 2024. 3

[33] Yue Liu, Shengfang Zhai, Mingzhe Du, Yulin Chen, Tri Cao, Hongcheng Gao, Cheng Wang, Xinfeng Li, Kun Wang, Junfeng Fang, et al. Guardreasoner-vl: Safeguarding vlms via reinforced reasoning. *arXiv preprint arXiv:2505.11049*, 2025. 2

[34] Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*, 2025. 3

[35] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *NeurIPS*, 2022. 3

[36] Ahmed Masry, Do Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *ACL*, 2022. 3

[37] Rameswar Panda, Jianming Zhang, Haoxiang Li, Joon-Young Lee, Xin Lu, and Amit K Roy-Chowdhury. Contemplating visual emotions: Understanding and overcoming dataset bias. In *ECCV*, pages 579–595, 2018. 6

[38] Kuan-Chuan Peng, Tsuhan Chen, Amir Sadovnik, and Andrew C Gallagher. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *CVPR*, pages 860–868, 2015. 6

[39] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *NeurIPS*, 36:53728–53741, 2023. 2

[40] Neel Rajani, Aryo Pradipta Gema, Seraphina Goldfarb-Tarrant, and Ivan Titov. Scalpel vs. hammer: Grpo amplifies existing capabilities, sft replaces them. *arXiv preprint arXiv:2507.10616*, 2025. 1

[41] Shyam Sundhar Ramesh, Yifan Hu, Iason Chaimalas, Viraj Mehta, Pier Giuseppe Sessa, Haitham Bou Ammar, and Ilija Bogunovic. Group robust preference optimization in reward-free rlhf. In *NeurIPS*, pages 37100–37137, 2024. 2, 3

[42] Maxime Robeyns and Laurence Aitchison. Improving llm-generated code quality with grpo. *arXiv preprint arXiv:2506.02211*, 2025. 2

[43] Xuankun Rong, Wenke Huang, Jian Liang, Jinhe Bi, Xun Xiao, Yiming Li, Bo Du, and Mang Ye. Backdoor cleaning without external guidance in mllm fine-tuning. *arXiv preprint arXiv:2505.16916*, 2025. 1

[44] Xuankun Rong, Wenke Huang, Tingfeng Wang, Daiguo Zhou, Bo Du, and Mang Ye. Safegrpo: Self-rewarded multimodal safety alignment via rule-governed policy optimization. *arXiv preprint arXiv:2511.12982*, 2025. 2

[45] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 2, 3

[46] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 2, 3, 6

[47] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, pages 8317–8326, 2019. 3

[48] Chengzhuo Tong, Ziyu Guo, Renrui Zhang, Wenyu Shan, Xinyu Wei, Zhenghao Xing, Hongsheng Li, and Pheng-Ann Heng. Delving into rl for image generation with cot: A study on dpo vs. grpo. *arXiv preprint arXiv:2505.17017*, 2025. 2

[49] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 3

[50] Cheng Wang, Yue Liu, Baolong Li, Duzhen Zhang, Zhongzhi Li, and Junfeng Fang. Safety in large reasoning models: A survey. *arXiv preprint arXiv:2504.17704*, 2025. 3

[51] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1, 2

[52] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 35:24824–24837, 2022. 3

[53] Hongxia Xie, Chu-Jun Peng, Yu-Wen Tseng, Hung-Jen Chen, Chan-Feng Hsu, Hong-Han Shuai, and Wen-Huang Cheng. Emovit: Revolutionizing emotion insights with visual instruction tuning. In *CVPR*, pages 26596–26605, 2024. 1, 3

[54] Bohao Xing, Zitong Yu, Xin Liu, Kaishen Yuan, Qilang Ye, Weicheng Xie, Huanjing Yue, Jingyu Yang, and Heikki Kälviäinen. Emo-llama: Enhancing facial emotion understanding with instruction tuning. *arXiv preprint arXiv:2408.11424*, 2024. 1, 3

[55] Shusheng Xu, Wei Fu, Jiaxuan Gao, Wenjie Ye, Weilin Liu, Zhiyu Mei, Guangju Wang, Chao Yu, and Yi Wu. Is dpo superior to ppo for llm alignment? a comprehensive study. *arXiv preprint arXiv:2404.10719*, 2024. 2

[56] Dingkang Yang, Zhaoyu Chen, Yuzheng Wang, Shunli Wang, Mingcheng Li, Siao Liu, Xiao Zhao, Shuai Huang, Zhiyan Dong, Peng Zhai, et al. Context de-confounded emotion recognition. In *CVPR*, pages 19005–19015, 2023. 1

[57] Jingyuan Yang, Qirui Huang, Tingting Ding, Dani Lischinski, Danny Cohen-Or, and Hui Huang. Emoset: A large-scale visual emotion dataset with rich attributes. In *ICCV*, pages 20383–20394, 2023. 6

[58] Qu Yang, Mang Ye, and Bo Du. Emollm: Multimodal emotional understanding meets large language models. *arXiv preprint arXiv:2406.16442*, 2024. 1, 3

[59] Zhicheng Yang, Zhijiang Guo, Yinya Huang, Xiaodan Liang, Yiwei Wang, and Jing Tang. Treerpo: Tree relative policy optimization. *arXiv preprint arXiv:2506.05183*, 2025. 2

[60] Huanjin Yao, Qixiang Yin, Jingyi Zhang, Min Yang, Yibo Wang, Wenhao Wu, Fei Su, Li Shen, Minghui Qiu, Dacheng Tao, et al. R1-sharevl: Incentivizing reasoning capability of multimodal large language models via share-grpo. *arXiv preprint arXiv:2505.16673*, 2025. 2

[61] Mang Ye, Xuankun Rong, Wenke Huang, Bo Du, Nenghai Yu, and Dacheng Tao. A survey of safety on large vision-language models: Attacks, defenses and evaluations. *arXiv preprint arXiv:2502.14881*, 2025. 1

[62] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78, 2014. 3

[63] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025. 2, 6

[64] Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. *arXiv preprint arXiv:2503.12937*, 2025. 2, 3

[65] Liyun Zhang, Zhaojie Luo, Shuqiong Wu, and Yuta Nakashima. Microemo: Time-sensitive multimodal emotion recognition with subtle clue dynamics in video dialogues. In *ACM MM Workshop*, pages 110–115, 2024. 1, 3

[66] Shenao Zhang, Sirui Zheng, Shuqi Ke, Zhihan Liu, Wanxin Jin, Jianbo Yuan, Yingxiang Yang, Hongxia Yang, and Zhaoran Wang. How can llm guide rl? a value-based approach. *arXiv preprint arXiv:2402.16181*, 2024. 3

[67] Hongjin Zhao, Zheyuan Liu, Yang Liu, Zhenyue Qin, Jiaxu Liu, and Tom Gedeon. Facephi: Lightweight multimodal large language model for facial landmark emotion recognition. In *ICLR Workshop*, 2024. 1, 3

[68] Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. Galore: Memory-efficient llm training by gradient low-rank projection. *arXiv preprint arXiv:2403.03507*, 2024. 3

[69] Jiaxing Zhao, Xihan Wei, and Liefeng Bo. R1-omni: Explainable omni-multimodal emotion recognition with reinforcement learning. *arXiv preprint arXiv:2503.05379*, 2025. 2, 3

[70] Guanghao Zhou, Panjia Qiu, Cen Chen, Jie Wang, Zheming Yang, Jian Xu, and Minghui Qiu. Reinforced mllm: A survey on rl-based reasoning in multimodal large language models. *arXiv preprint arXiv:2504.21277*, 2025. 2, 3