DSPO: DIVERSITY-AWARE SUBJECTIVE POLICY OPTIMIZATION FOR ROBUST EMOTIONAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement Learning has significantly advanced the complex reasoning capabilities of Multimodal Large Language Models (MLLMs). However, prevailing RL algorithms, such as Group Relative Policy Optimization (GRPO), suffer a severe failure in emotion reasoning tasks. These methods heavily rely on deterministic hard-label supervision and point-wise isolated evaluation, creating a fundamental gap with the inherently subjective and continuously distributed nature of human emotions. Furthermore, unlike explicit physical objects, emotional states are deeply implicit within visual cues. This abstract nature exacerbates visual hallucinations in MLLMs, leading to plausible yet ungrounded emotional evidence. To address these limitations, we propose **Diversity-Aware Subjective Policy Optimization (DSPO)**, a reinforcement learning framework that jointly promotes subjective affective coverage and visual grounding. First, we construct a context-grounded emotional distribution prior in the VAD space by combining the lexical prior of the annotated emotion with image-specific contextual information. Based on this prior, we introduce a Distribution-Aligned Emotional Diversity Reward (DEDR), which measures the leave-one-out marginal contribution of each candidate emotion within a rollout. DEDR rewards candidates whose inclusion brings the predicted affective set closer to the context-grounded prior, thereby preserving plausible subjective interpretations without encouraging unconstrained dispersion. We further develop Counterfactual Visual Intervention Gating (CVIG), which masks the visual region highlighted in the reasoning process and uses the resulting candidate-wise probability changes to reduce the weights of interpretations unsupported by visual evidence. Extensive experiments demonstrate that DSPO achieves state-of-the-art performance across multiple public benchmarks, especially on the cross-domain performance, *i.e.,* improving +10.8% on average cross-domain accuracy than EMO-R3.[1]

1 INTRODUCTION

Multimodal Large Language Models (MLLMs) have recently demonstrated remarkable proficiency in general-oriented perception and reasoning tasks (Yang et al., 2026; Liu et al., 2025; Yue et al., 2025). However, they frequently failed when applied to emotional computing and human-centered reasoning (Xie et al., 2024; Zhang et al., 2025; Dingdong et al.). Unlike object-centric tasks (Yao et al., 2026a; Zhang et al., 2026; Gu et al., 2026), where ground truths are deterministic and visually explicit, emotion states are inherently latent, highly subjective, and deeply embedded within nuanced multimodal contexts. Evaluating emotion states requires models to go beyond superficial pattern recognition to capture micro-level expressions, complex social dynamics, and subtle psychological signals (Qin et al., 2026; Wang et al., 2026; Yuan et al., 2026). Consequently, bridging the semantic gap between explicit visual elements and implicit human emotions remains a grand challenge for current MLLMs (Guo et al., 2025; Chen et al., 2026a; Yao et al., 2026b).

Recently, Reinforcement Learning (RL), particularly Group Relative Policy Optimization (GRPO) (Shao et al., 2024), has emerged as a promising paradigm to elicit advanced reasoning capabilities in MLLMs (Chaubey et al.; Ge et al., 2026). Despite its success in deterministic tasks like mathematical reasoning, applying GRPO directly to emotion reasoning reveals two intrinsic

[1]Code will be released in the final version of the paper.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

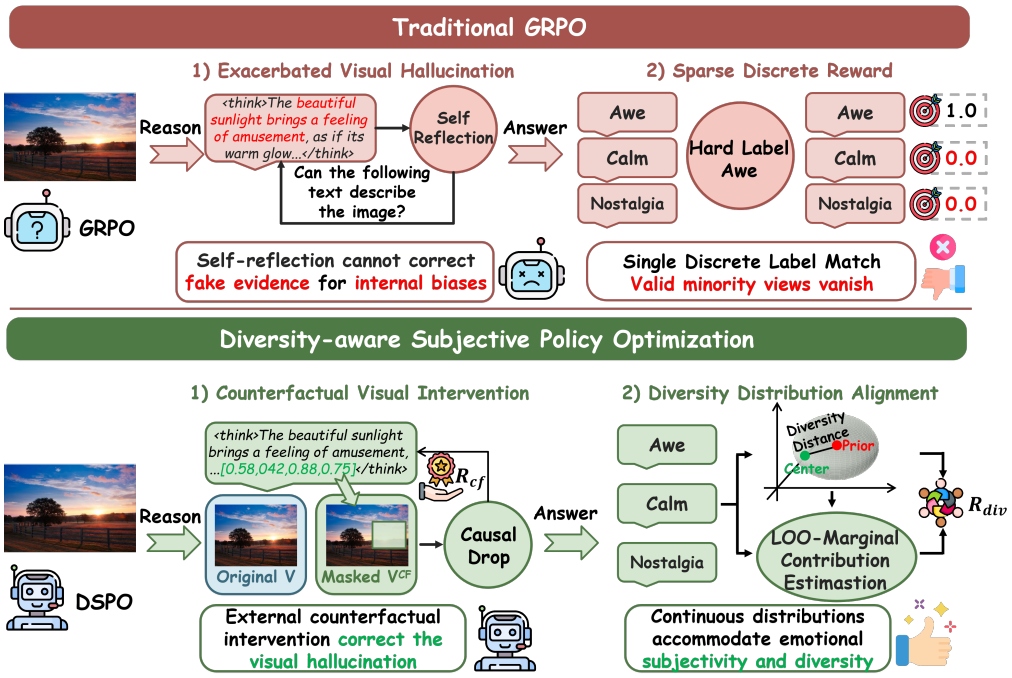

Figure 1: Comparison between traditional GRPO and our DSPO. 1) Traditional self-reflection mechanisms struggle to correct the internal biases of MLLMs. Counterfactual visual intervention mitigates hallucinations about visual evidence. 2) Diversity Distribution Reward addresses the issue where the sparse, discrete rewards of GRPO struggle to accommodate sentiment-close answers.

limitations. 1) **Exacerbated Visual Hallucination.** When inferring implicit affective cues, the tendency of MLLMs to generate visual hallucinations is significantly exacerbated. Specifically, models frequently generate non-existent visual evidence to forcibly justify a plausible emotional conclusion. More critically, in the absence of objective physical anchoring, existing methods (Fang et al., 2026) attempt to correct evidence-image inconsistencies through self-reflection mechanisms, which lacks external verification and easily fall prey to confirmation bias within MLLMs. This closed-loop self-justification not only amplifies reasoning errors but also reinforces hallucinated reasoning trajectories. Ultimately, this causes MLLMs to merely adopt shortcut learning to accommodate target labels without acquiring causal affective reasoning capabilities, thereby severely compromising their generalization and robustness in open-world scenarios. 2) **Sparse Discrete Reward.** Existing RL-based methods rely on discrete emotion labels as reward signals. This rigid matching mechanism fundamentally conflicts with the continuous and distributed nature of human emotion, where diverse subjective interpretations could naturally coexist. Penalizing valid minority perspectives inevitably forces the policy into mode collapse, causing it to merely fit the label distribution of the given dataset rather than learning the true emotional semantic space.

To address these limitations, we propose Distribution-level Subjective Policy Optimization (DSPO), a novel reinforcement learning framework specifically tailored to align MLLMs with both human subjectivity and objective visual causality. DSPO constructs a context-grounded emotional distribution prior as an enhanced optimization signal by combining the lexical prior of the annotated emotion with image-specific contextual information. Then DSPO introduces a Distribution-Aligned Emotional Diversity Reward, which calculates the leave-one-out marginal contribution of each candidate emotion within a rollout and rewards candidates who makes the predicted entire distribution closer to the prior, thereby preserving plausible subjective interpretations without encouraging unconstrained dispersion. Besides, to ensure this empirical distribution is not contaminated by hallucinated reasoning, we introduce a Counterfactual Visual Intervention Gating (CVIG). By enforcing explicit physical grounding and computing the causal probability drop under latent visual masking, CVIG strictly penalizes visual hallucinations and assigns a weight to each candidate emotion. Ultimately, DSPO harmonizes the subjective diversity of emotional expression with the causal validity of visual evidence, significantly enhancing the emotional reasoning capabilities of MLLMs. In summary, our main contributions are as follows:

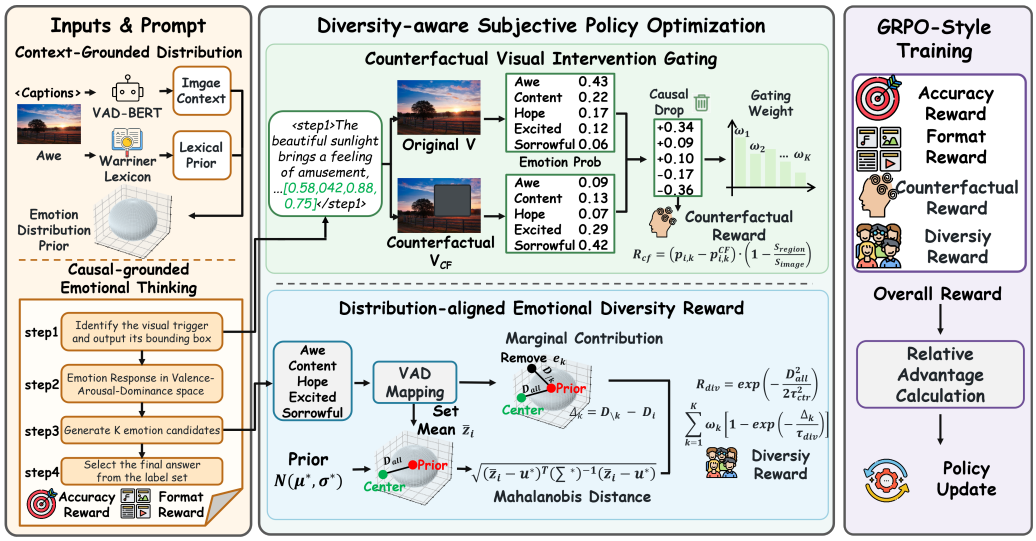

Figure 2: Overview of DSPO framework. The left part presents the construction of the emotional distribution prior and the input prompt. The middle part show two crucial counterfactual visual intervention gating and distribution-aligned emotional diversity reward modules. Finally, DSPO are jointly optimized with the original Format and Accuracy rewards under the GRPO framework.

• We propose Diversity-aware Subjective Policy Optimization (DSPO), a novel RL framework that introduces a continuous emotion distribution prior as an enhanced optimization signal, addressing the sparse reward by discrete hard-labels in existing RL training for emotion reasoning tasks.

• We introduce a Distribution-Aligned Emotional Diversity Reward that calculates the subjective diversity of all generated candidate emotions within a rollout, naturally tolerating human emotional subjectivity without mode collapse. Besides, we design a Counterfactual Visual Intervention Gating, which verifies the validity of visual evidence by evaluating drops of causal probability under visual masking, alleviating exacerbated visual hallucinations in implicit affective reasoning.

• Extensive experiments demonstrate that DSPO achieves state-of-the-art performance on multiple emotion reasoning benchmarks, especially on the cross-domain performance, *i.e.,* improving +10.8% on average cross-domain accuracy than EMO-R3.

2 DSPO: Diversity-aware Subjective Policy Optimization

2.1 Preliminary

We evaluate the emotional reasoning ability using an image emotion recognition task. It is crucial to design an explicitly guided instruction prompt for the internal thinking phase of MLLMs. Generic Chain-of-Thought (CoT) prompts (*e.g.*, "Please think step by step.") inherently lack task-specific designs. Consequently, they fail to establish a causal mapping between low-level visual cues and high-level abstract emotional conclusions. Such unconstrained exploration struggles to elicit reasoning trajectories that align with human affective cognition, and instead frequently exacerbates factual drift and visual hallucinations. To overcome this, we design a Cause-grounded Emotional Thinking.[2] Specifically, we first guide the MLLM identify key visual evidence and output its location by a bounding box. Based on this visual evidence, we require the MLLM to output emotional responses. Subsequently, unlike previous methods, we do not directly require the MLLM to output a single emotion category. Instead, we allow the MLLM to first generate K potential emotion candidates and then combine them to select the best match from the given label set. Such thinking mode enhances the mining of visual evidence through physical anchoring, while increasing the diversity of emotion reasoning by generating a candidate set of emotions. Overall, we adopt a GRPO-style framework. Given an image-prompt input pair $(\mathcal{V}, Q)$, the MLLM $\pi_{\theta_{old}}$ receives the input pair and

[2]The details of the thinking template are shown in the Appendix.

samples a group of G distinct rollouts, denoted as $\mathcal{O} = \{o_1, o_2, \ldots, o_G\}$. Finally, the MLLM is optimized by maximizing the following objective function:

$$\mathcal{J}(\theta) = \mathbb{E}_{(\mathcal{V},Q)} E_{\mathcal{O}\sim\pi_{\theta_{old}}} \left[\frac{1}{G} \sum_{i=1}^{G} \Big(\min \left[\rho_i(\theta)\hat{A}_i,\ \text{clip}\left(\rho_i(\theta), 1-\epsilon, 1+\epsilon\right)\hat{A}_i \right] \right] - \beta \mathbb{D}_{KL}(\pi_\theta \,||\, \pi_{ref}) \tag{1}$$

where ϵ, β are clipping and divergence penalty hyper-parameters and the importance sampling ratio $\rho_i(\theta)$ is defined as $\rho_i(\theta) = \frac{\pi_\theta(o_i|(\mathcal{V},Q))}{\pi_{\theta_{old}}(o_i|(\mathcal{V},Q))}$.

2.2 Context-Grounded Emotional Distribution Prior

To provide a reliable prior for continuous emotion distributions. We design a context-grounded emotional distribution prior construction pipeline. A natural initial intuition is to leverage the widely used Warriner lexicon (Warriner et al., 2013) to map discrete emotion labels into continuous VAD-Gaussian distributions. However, this approach completely ignores the nuanced affective variations induced by diverse visual scenes in the real world. Therefore, we further introduce a dynamic context driven by the specific image content. Specifically, for the lexical prior, we query the Warriner lexicon to retrieve the baseline distribution parameters representing human consensus. For the dynamic context, we first process the image through an advanced MLLM Mimo-v2 (Xiao et al., 2026) to generate a detailed caption. This caption is subsequently fed into a pre-trained sentence-level VAD regression model (Buechel & Hahn, 2017) to extract context-specific distribution parameters:

$$\boldsymbol{\mu}_{sta}, \boldsymbol{\sigma}_{sta} = \Phi_{lex}(y), \mathcal{C} = \text{CapGen}(\mathcal{V}), \boldsymbol{\mu}_{ctx} = \text{VADReg}(\mathcal{C}), \tag{2}$$

where y is the ground-truth emotion category and Φ_{lex}, CapGen, VADReg denote Warriner lexicon, caption generator, and VAD regression model, respectively. Besides, to quantify the ambiguity between the image context and lexical prior, we calculate the contextual divergence vectors:

$$\boldsymbol{\sigma}_{ctx} = \text{Var}(\boldsymbol{\mu}_{sta}, \boldsymbol{\mu}_{ctx}), \tag{3}$$

where Var denotes the variance calculation. Finally, we derive the ultimate ground-truth distribution label by aggregating these two sets of parameters via weighted fusion:

$$\boldsymbol{\mu}^* = \frac{\boldsymbol{\mu}_{sta} + \boldsymbol{\mu}_{ctx}}{2}, \boldsymbol{\sigma}^* = \frac{\boldsymbol{\sigma}_{sta} + \boldsymbol{\sigma}_{ctx}}{2}, \tag{4}$$

The Gaussian distribution $\mathcal{Z} \sim \mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\sigma}^*)$ represents the ground-truth human emotion distribution, serving as the gold standard for calculating emotion diversity.

2.3 Counterfactual Visual Intervention Gating

Prior works (Fang et al., 2026) attempt to mitigate hallucinations through reflective emotional rewards. However, these approaches typically restrict themselves to prompt-level self-correction over existing outputs, leaving the internal self-bias of MLLMs fundamentally unaddressed. Moreover, this pattern encourages the MLLM to generate fake emotional cues to justify a correct emotion conclusion. We analyze that the core question is whether the generated textual claims actually point to a visually present region. Such causal attribution regarding the reasoning chain is independent of external emotional labels. To this end, we introduce a Counterfactual Visual Intervention Gating (CVIG) module. Concretely, we generate a counterfactual image by masking the pointed visual regions and measure the shift in prediction probabilities for each candidate emotion, which quantifies the consistency between the specified visual region and each emotional interpretation. Specifically, we first extract the emotional cue text and its associated bounding box coordinates from the output:

$$(t_i^1, [x_i^{min}, y_i^{min}, x_i^{max}, y_i^{max}]) = \Phi_{regex}\big(o_i, \texttt{<step1>}(.*?)\texttt{</step1>}\big), \tag{5}$$

where t_i^1 represents the text segment for the visual evidence and Φ_{regex} denotes the regex extractor. Subsequently, we map the extracted bounding box coordinates to the indices of the visual encoder:

$$\hat{x}_i^{\min} = \lfloor x_i^{\min} \cdot W \rfloor, \quad \hat{x}_i^{\max} = \lceil x_i^{\max} \cdot W \rceil, \quad \hat{y}_i^{\min} = \lfloor y_i^{\min} \cdot H \rfloor, \quad \hat{y}_i^{\max} = \lceil y_i^{\max} \cdot H \rceil, \tag{6}$$

$$\mathcal{T}_i = \left\{ (r, c) \mid \hat{y}_i^{\min} \le r < \hat{y}_i^{\max},\ \hat{x}_i^{\min} \le c < \hat{x}_i^{\max} \right\}, \tag{7}$$

where H, W is the size of the image. For visual features $\mathcal{V}_i \in \mathbb{R}^{H\times W\times D}$, we obtain the counterfactual image by constructing a mask matrix $M_i \in \mathbb{R}^{H\times W}$:

$$M_i(j,k) = \begin{cases} 0, & (j,k) \in \mathcal{T}_i, \\ 1, & \text{otherwise}, \end{cases} \quad \text{and} \quad \mathcal{V}_i^{CF} = \mathcal{V}_i \odot M_i, \tag{8}$$

where $\odot$ denotes element-wise multiplication. We feed the original and counterfactual image features $\mathcal{V}_i$ and $\mathcal{V}_i^{CF}$ into the subsequent decoder and calculate the predicted probabilities for each of the K candidate emotion categories extracted from the <step3> list:

$$p_{i,k} = P\big(e_i^{(k)} \mid \mathcal{V}_i, \mathcal{Q}, t_i\big), \qquad p_{i,k}^{CF} = P\big(e_i^{(k)} \mid \mathcal{V}_i^{CF}, \mathcal{Q}, t_i\big), \quad k = 1, 2, \ldots, K, \tag{9}$$

where $e_i^{(k)}$ denotes the k-th emotion candidate from the <step3> list. We then compute the causal drop for each candidate as the relative decline in its prediction probability under masking, with an area penalty to prevent the model from selecting the entire image. Finally, we compute the total drop of all candidates as the counterfactual reward $\mathcal{R}_{cf}$:

$$\Delta P_{i,k} = \big(p_{i,k} - p_{i,k}^{CF}\big) \cdot \left(1 - \frac{S_{\text{region}}}{S_{\text{image}}}\right), \quad \mathcal{R}_{cf,i} = \frac{1}{K}\sum_{k=1}^{K} \Delta P_{i,k} \tag{10}$$

where S_{region} and S_{image} denote the area of the anchored region and the entire image, respectively. Besides, we compute the normalized counterfactual weights across all K candidates:

$$w_{i,k} = \frac{\exp(\Delta P_{i,k}/\tau)}{\sum_{j=1}^{K} \exp(\Delta P_{i,j}/\tau)}, \qquad \tau > 0, \tag{11}$$

where τ is the temperature parameter controlling the sharpness of the distribution. These gating weights are directly applied to the diversity reward to down-weight hallucinated candidates.

2.4 Distribution-aligned Emotional Diversity Reward

To address the limitation of single discrete label and align the continuous and subjective human distribution, we introduce a distribution-aligned emotional diversity reward that encourages the model to produce a distribution of plausible emotional interpretations based on the constructed prior.

Specifically, we first extract the text segment of emotion response and K candidate emotions within each rollout:

$$t_i^2, \left\{e_i^{(1)}, e_i^{(2)}, \ldots, e_i^{(K)}\right\} = \Phi_{regex}\big(o_i, [\texttt{<step2>(.*?)</step2>}, \texttt{<step3>(.*?)</step3>}]\big), \tag{12}$$

then each emotion word $e_i^{(k)}$ is mapped to a VAD vector using the Warriner lexicon:

$$\mathbf{z}_i^{(k)} = \Phi_{\text{lex}}\big(e_i^{(k)}\big) \in \mathbb{R}^3, \quad k = 1, 2, \ldots, K, \tag{13}$$

where the three dimensions correspond to Valence, Arousal, and Dominance, respectively. For out-of-vocabulary words, we apply a stemmer-based fallback or nearest-neighbor lookup to ensure robust coverage. Subsequently, to capture the subjective diversity of emotional interpretations beyond mere dispersion, we adopt a Leave-One-Out (LOO) marginal contribution estimation. We first compute the VAD-vector centroid of K candidate emotions. The diversity baseline is defined as the Mahalanobis distance from the centroid to our constructed prior distribution:

$$\bar{\mathbf{z}}_i = \frac{1}{K}\sum_{k=1}^{K} \mathbf{z}_i^{(k)}, \quad D_{i,all} = \sqrt{(\bar{\mathbf{z}}_i - \boldsymbol{\mu}^*)^\top (\boldsymbol{\Sigma}^*)^{-1} (\bar{\mathbf{z}}_i - \boldsymbol{\mu}^*)}, \tag{14}$$

where $\boldsymbol{\Sigma}^* = \begin{vmatrix} \sigma_V & & \\ & \sigma_A & \\ & & \sigma_D \end{vmatrix}$ is the covariance matrix. Subsequently, for each candidate, we remove it and recompute the distance using the remaining $K-1$ candidates:

$$D_{i,\setminus k} = \sqrt{(\bar{\mathbf{z}}_{i,\setminus k} - \boldsymbol{\mu}^*)^\top (\boldsymbol{\Sigma}^*)^{-1} (\bar{\mathbf{z}}_{i,\setminus k} - \boldsymbol{\mu}^*)}, \quad \bar{\mathbf{z}}_{i,\setminus k} = \frac{1}{K-1}\sum_{j\neq k} \mathbf{z}_i^{(j)}, \tag{15}$$

Table 1: Comparison with GRPO variants and SOTA methods across in-domain and out-of-domain settings. The best and suboptimal results are highlighted in bold and underline, respectively.

Methods	Rollout	EmoSet^I	Emotion6	WebEmo	Emotion6^I	EmoSet	WebEmo	$\mathcal{A}^I$	$\mathcal{A}^O$	$\mathcal{A}$
LLaVA-1.5-7B										
Zero-shot	-	52.77	48.32	25.56	48.32	52.77	25.56	50.55	38.05	42.22
SFT	-	56.04	54.21	42.39	54.21	56.04	42.39	55.13	48.76	50.88
Qwen2.5-VL-3B-Instruct										
Zero-shot	-	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 (2024)	4	74.60	60.10	49.50	<u>70.88</u>	59.90	44.85	72.74	53.59	59.97
DAPO (2026)		68.99	56.90	49.80	68.56	59.95	<u>45.50</u>	68.78	53.04	58.28
EMO-R3 (2026)		**75.50**	<u>60.44</u>	<u>50.45</u>	70.71	<u>60.70</u>	45.20	<u>73.10</u>	<u>54.20</u>	<u>60.50</u>
Ours		<u>75.30</u>	**67.31**	**54.54**	**71.20**	**69.80**	**51.80**	**73.25**	**60.86**	**64.99**
GRPO (2024)	8	75.45	57.91	49.40	69.87	60.30	42.05	72.66	52.42	59.16
DAPO (2026)		70.21	55.72	48.80	62.39	58.05	<u>46.30</u>	66.30	52.22	56.91
EMO-R3 (2026)		<u>76.40</u>	<u>59.26</u>	<u>49.70</u>	<u>71.72</u>	<u>61.80</u>	43.65	<u>74.06</u>	<u>53.60</u>	<u>60.42</u>
Ours		**76.65**	**66.87**	**54.00**	**71.80**	**67.40**	**49.25**	**74.23**	**59.38**	**64.33**

the marginal contribution of $e_i^{(k)}$ is defined as the increase in distance. A positive increase indicates that the candidate carries a unique subjective perspective that brings the group closer to the human prior. Then we design the diversity reward $\mathcal{R}_{div}$ by considering the centroid distance and the sum of the marginal contribution of all candidates:

$$\Delta_{i,k} = D_{i,\setminus k} - D_i, \quad \mathcal{R}_{div,i} = \exp\left(-\frac{D_{i,all}^2}{2\tau_{\text{ctr}}^2}\right) \sum_{k=1}^{K} w_{i,k} \left[1 - \exp\left(-\frac{\Delta_{i,k}}{\tau_{\text{div}}}\right)\right], \tag{16}$$

where $\tau_{\text{ctr}}, \tau_{\text{div}}$ are temperature coefficient. By comprehensively considering the accuracy of the central distribution and the internal emotional diversity of the candidate set, we guide the MLLM to achieve interpretable emotional reasoning that aligns with human subjective diversity.

2.5 Overall Reward and Training

Besides the above two designed rewards $\mathcal{R}_{cf}$ and $\mathcal{R}_{div}$, following the traditional GRPO training, we define two general rewards to guide the optimization of the structured emotional reasoning. We first define the format reward $\mathcal{R}_{fmt}$ to measure whether the generated reasoning text adheres to the predefined structure. Specifically, it checks whether each reasoning step corresponds to the expected stage `<stepi>...</stepi>` and whether the bounding box, the emotion list, and the chosen option are correctly enclosed in `\bboxed{}`, `\list{}`, and `\boxed{}`:

$$\mathcal{R}_{\text{format}} = \begin{cases} 1, & \text{if } \texttt{\bboxed\{\}}, \texttt{\list\{\}}, \texttt{\boxed\{\}} \text{ formats are correct,} \\ 0, & \text{otherwise.} \end{cases} \tag{17}$$

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

$$\mathcal{R}_{\text{acc}} = \begin{cases} 1, & \text{if } \hat{\mathcal{E}} = \mathcal{E}^*, \\ 0, & \text{otherwise.} \end{cases} \tag{18}$$

Finally, the total reward of the DSPO training is calculated by the weighted sum of all four rewards:

$$\mathcal{R}_{total} = \lambda_{acc} \cdot \mathcal{R}_{acc} + \lambda_{fmt} \cdot \mathcal{R}_{fmt} + \lambda_{cf} \cdot \mathcal{R}_{cf} + \lambda_{div} \cdot \mathcal{R}_{div}, \tag{19}$$

where $\lambda_{acc}, \lambda_{fmt}, \lambda_{cf}, \lambda_{div}$ are hyper-parameters to control the balance between different rewards.

3 Result and Discussion

3.1 Main Results

As shown in Table 1, we first observe that DSPO achieves the highest overall accuracy of in- and out-of- domain under both $G = 4/8$, *i.e.,* +46.7%/+8.7% improvements than SFT/EMO-R3 when $G = 8$. This demonstrates the superior performance of DSPO across various scenarios involving emotion understanding. Besides, we observe that the improvements are particularly pronounced

under out-of-domain evaluation, *i.e.,* +13.6%/+12.3% improvements than GRPO/EMO-R3 when $G = 8$. Such significant gains indicate that DSPO improves more than in-domain label fitting. The distribution-aligned diversity reward preserves plausible neighboring affective interpretations in the continuous VAD space, while CVIG suppresses candidates unsupported by visual evidence, jointly reducing label-specific shortcut learning. Meanwhile, DSPO maintains a competitive in-domain performance and the in-domain accuracy under two rollout settings is slightly higher than EMO-R3, which indicates that diversity reward learning does not come at the expense of fitting in-domain labels and instead enhances the affective semantic space upon that foundation.

We also evaluate on four distribution-based metrics.[3] As shown in Table 2, we first observe that DSPO achieves the best performance on three relative metrics, *i.e.,* KL, JS, and En-MAE. This indicates that DSPO could generate sampling distributions that are closer to the true human distribution. Furthermore, we observe that adding an entropy-based reward to GRPO improves H^{model} but leads to a significant decline across three relative metrics. This demonstrates that the unconstrained diversity generation fails to enhance emotional reasoning capabilities. In contrast, DSPO calculates a diversity reward aligned with the human distribution, enabling the model to achieve a balance between accurate emotional reasoning and diverse emotional generation.

Table 2: The results for distribution-based evaluation.

Methods	KL↓	JS↓	H^{model} ↑	En-MAE ↓
SFT	1.940	0.457	0.395	0.397
GRPO	0.702	0.260	0.581	0.298
GRPO+$\mathcal{R}_{en}$	0.884	0.295	**0.763**	0.329
DAPO	1.059	0.328	0.524	0.348
EMO-R3	0.650	0.244	0.623	0.274
DSPO	**0.372**	**0.169**	0.696	**0.209**

3.2 Ablation Studies

Discussion on proposed modules. Table 3 explores the contributions of CVIG and DEDR. First, we observe that using CVIG alone primarily improves in-domain accuracy and even degrades on the out-of-domain WebEmo dataset. CVIG enhances the authenticity and accuracy of emotional reasoning by filtering out spurious visual evidence. Besides, using DEDR alone significantly improves out-of-domain performance. By designing a diversity reward that guides the MLLM to learn continuous emotion distributions rather than only fitting the label distribution of training source via hard labels, DEDR substantially enhances the robustness. Finally, the synergy between the CVIG and DEDR modules further enhances overall performance, thereby enabling a more comprehensive understanding of emotion.

Table 3: The ablation study for CVIG and DEDR modules.

CVIG	DEDR	EmoSet^I	Emotion6	WebEmo	$\mathcal{A}$
×	×	75.45	57.91	49.40	60.92
✓	×	76.10	58.85	48.91	61.29
×	✓	75.60	64.25	52.60	64.15
✓	✓	**76.65**	**66.87**	**54.00**	**65.84**

Impact of Candidate Emotion Number. Table 4 explore the impact of different numbers of candidate emotions within each rollout. We first observe a significant performance drop when $K = 1$. A single candidate is difficult to adequately represent the ambiguity of human emotions and fit the distribution prior. A moderate candidate set allows the model to cover multiple plausible regions of the target VAD distribution and provides more informative marginal-contribution estimates. Nevertheless, model performance begins to decline as increasing $K = 7$, suggesting that an excessively large candidate set introduces redundant grounded emotions and adds noise to distribution matching. Finally, we set $K = 5$ as an effective balance between subjective coverage and candidate reliability.

Table 4: The ablation study for the number of candidate emotions.

Top-K	EmoSet^I	Emotion6	WebEmo	$\mathcal{A}$
1	72.35	59.20	50.15	60.57
3	75.90	62.74	51.20	63.28
5	**76.65**	**66.87**	**54.00**	**65.84**
7	74.45	60.81	50.99	62.08

[3]Settings of dataset, metrics, and implementation details are shown in the Appendix.

Table 5: The discussion for diversity computation.

Setting	EmoSet^I	Emotion6	WebEmo	$\mathcal{A}$
w/o DEDR	76.10	58.85	48.91	61.29
Center Distance	76.30	60.01	48.75	61.69
Pairwise Distance	74.74	58.40	47.89	60.34
LOO-Margin Distance	75.09	65.49	52.60	64.39
Combined Distance	**76.65**	**66.87**	**54.00**	**65.84**

Discussion on diversity computation. Table 5 compares different diversity computation strategies, *i.e.,* 1) Center Distance: Only calculate the distance between predicted entire distribution and prior. 2) Pairwise Distance: Calculate the distances between all pairs of candidate emotions. 3) LOO-Margin Distance: Only calculate the leave-one-out marginal contribution of each candidate emotion. We first observe that Pairwise Distance significantly degrades overall performance, indicating that focusing solely on in-set diversity without constraining it close to the prior is insufficient. Besides, considering center and LOO-Margin distance separately both fail to achieve the best performance, which indicates that we need to comprehensively consider both the accuracy of the entire set within the emotional semantic space and the diversity within the set.

Table 6: The discussion on CVIG module.

Setting	EmoSet^I	Emotion6	WebEmo	$\mathcal{A}$
w/o CVIG	75.60	64.25	52.60	64.15
Bounding Box				
Random Box	67.80	57.30	46.59	57.23
w/o Area Penalty	71.55	60.89	49.10	60.51
Intervention				
Mean Replace	72.11	60.60	49.70	60.80
Gaussian Noise	75.29	64.50	52.89	64.23
w/ CVIG	**76.65**	**66.87**	**54.00**	**65.84**

Discussion on CVIG module. Table 6 explores the effects of the bounding box and intervention settings in CVIG module. We first observe that using random box causes a substantial performance drop. This is due to the removal of emotion-related regions. Besides, removing the area penalty also degrades the average accuracy. Without this penalty, the model may prefer excessively large bounding boxes containing both relevant and irrelevant content, resulting in an imprecise counterfactual intervention. We further compare different intervention strategies for the selected regions. We observe that both gaussian noise and mean replacement reduce the performance. Gaussian noise introduces additional visual redundancy, and mean replacement does not completely remove the semantic information of the selected region. In contrast, zero masking provides a cleaner intervention by explicitly suppressing the selected visual features, thereby producing a clearer difference between the original and counterfactual predictions.

Impact of Reward Hyper-parameters. Fig. 3 investigates the sensitivity to λ_{cf} and λ_{div}. We first observe that increasing λ_{cf} from 0.01 to 0.1 mainly improves the in-domain performance. $\mathcal{R}_{cf}$ provides a grounding signal that suppresses fabricated emotional evidence. However, increasing λ_{cf} to 0.2 leads to a clear performance degradation. Overemphasizing visual consistency may amplify localization noise, suppress valid but subtle emotional cues. Besides, increasing λ_{div} from 0.01 to 0.3 mainly brings a significant improvement on out-of-domain performance. By encouraging the model to cover multiple plausible emotional interpretations around the human affective distribution, $\mathcal{R}_{div}$ reduces over-reliance on a single hard label and improves the robustness. However, the performance decreases on both in-domain and out-of-domain when increasing λ_{div} to 0.5, which indicates that overemphasizing emotional diversity will affect basic emotional reasoning abilities.

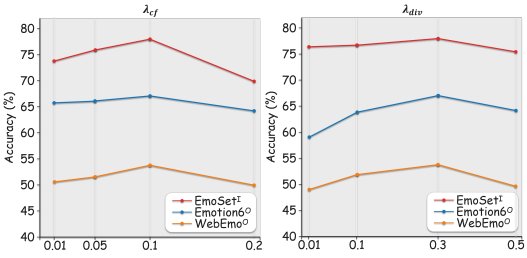

Figure 3: Impact for different reward weights.

3.3 Efficiency Analysis

Considering that we introduce additional modules, we conduct an efficiency analysis on the training process under the 8-rollout setting on EmoSet dataset. As shown in Fig. 5, we observe that although our model introduces a certain amount of extra computation overhead, it does not bring about a significant improvement in training time. Compared with EMO-R3 (Fang et al., 2026), our model

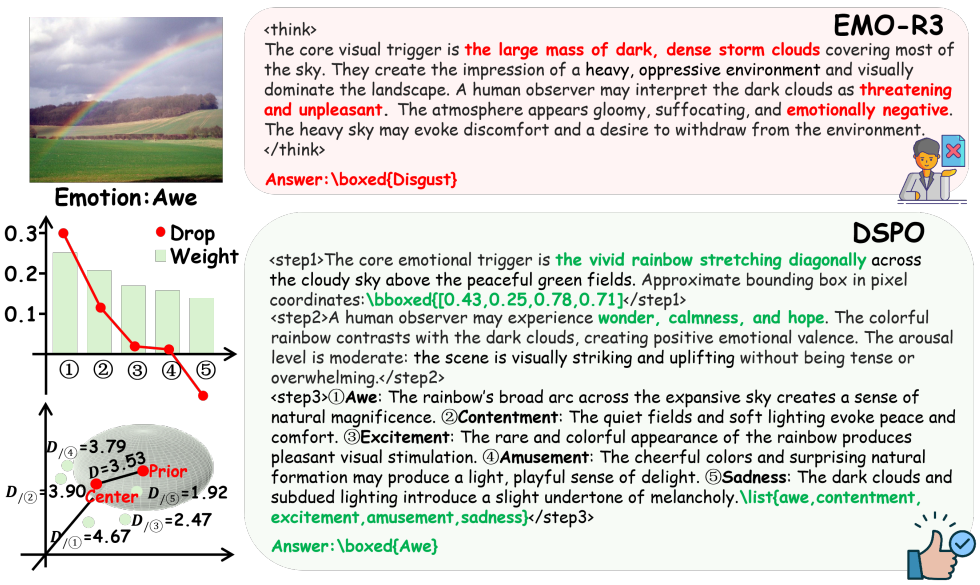

Figure 4: Case study between the most powerful method EMO-R3 and DSPO on the EmoSet dataset.

improves out-of-domain accuracy on the Emotion6 dataset by 12.8% while consuming only 27% more time. Moreover, the proposed gating and diversity module are both removed during the inference process, so that it requires no additional inference-time cost. Therefore, in real usage and evaluation, our model achieves better performance while maintaining high computational efficiency.

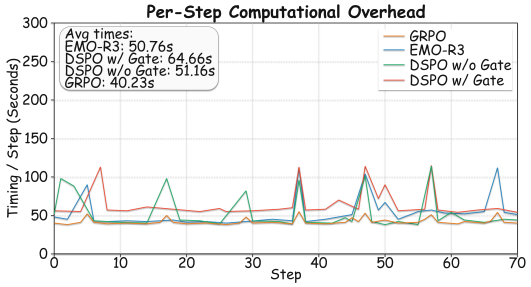

Figure 5: Efficiency analysis visualization.

3.4 Case Study

We present a case study to compare between baseline EMO-R3 (Fang et al., 2026) and DSPO. As shown in Fig. 4, we first observe that EMO-R3 incorrectly identifies the image as 'disgust', while DSPO correctly identifies it as 'awe'. Furthermore, we analyze that the reason is that EMO-R3 erroneously localizes visual evidence as dark clouds and hallucinates a 'heavy, oppressive environment'. Instead, from the weight distribution from CVIG, we find that DSPO detects this hallucination through counterfactual intervention and assigns the lowest weight to 'sadness'. Besides, DSPO could estimate human-aligned diversity. Specifically, removing 'sadness' reduces the overall distance, whereas removing 'awe' increases it. Overall, DSPO learns within a continuous emotion space rather than simply fitting the label distribution of the dataset as previous methods do.

4 Conclusion

In this paper, we introduce Diversity-Aware Subjective Policy Optimization (DSPO) for robust emotion reasoning, which addresses two limitations of conventional RL learning: the sparse discrete reward and exacerbated visual hallucinations. Specifically, to achieve continuous emotion supervision, we first construct a context-grounded emotional distribution prior and propose a Distribution-Aligned Emotional Diversity Reward that evaluates the marginal contribution of each candidate emotion to align the prior. Besides, to alleviate hallucinations when searching for visual evidence, we further introduce Counterfactual Visual Intervention Gating, which estimates candidate-wise causal drop through counterfactual masking. Extensive experiments demonstrate that DSPO consistently improves overall accuracy and delivers particularly strong cross-domain generalization.

REFERENCES

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.

Sven Buechel and Udo Hahn. Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 578–585, 2017.

Ashutosh Chaubey, Jiacheng Pang, Maksim Siniukov, and Mohammad Soleymani. Avere: Improving audiovisual emotion reasoning with preference optimization. In *The Fourteenth International Conference on Learning Representations*.

Weidong Chen, Cheng Ye, Zhendong Mao, Peipei Song, Xinyan Liu, Lei Zhang, Xiaojun Chang, and Yongdong Zhang. Face-net: Factual calibration and emotion augmentation for retrieval-enhanced emotional video captioning. *arXiv preprint arXiv:2603.17455*, 2026a.

Weidong Chen, Cheng Ye, Peipei Song, Lei Zhang, Yongdong Zhang, and Zhendong Mao. Subjective-objective emotion correlated generation network for subjective video captioning. *IEEE Transactions on Image Processing*, 2026b.

Zebang Cheng, Yuxiang Lin, Zhaoru Chen, Xiang Li, Shuyi Mao, Fan Zhang, Daijun Ding, Bowen Zhang, and Xiaojiang Peng. Semi-supervised multimodal emotion recognition with expression mae. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 9436–9440, 2023.

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

Zebang Cheng, Shuimu Chen, Boxue Yang, Yuanshen Guan, Jingyi Chen, Zheng Lian, Xiaojiang Peng, Fei Ma, LaiZhong Cui, and Qi Tian. Omniopsd: Rationale-privileged on-policy self-distillation for affective computing. *arXiv preprint arXiv:2606.15920*, 2026.

WANG Dingdong, LIU Shujie, M Meng Helen, et al. Emotionthinker: Prosody-aware reinforcement learning for explainable speech emotion reasoning. In *The Fourteenth International Conference on Learning Representations*.

Yiyang Fang, Wenke Huang, Pei Fu, Yihao Yang, Kehua Su, Zhenbo Luo, Jian Luan, and Mang Ye. Emo-r3: Reflective reinforcement learning for emotional reasoning in multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 745–755, 2026.

Shiran Ge, Chenyi Huang, Yuang Ai, Qihang Fan, Huaibo Huang, and Ran He. Expand and prune: Maximizing trajectory diversity for effective grpo in generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41913–41922, 2026.

Jiawei Gu, Yunzhuo Hao, Huichen Wang, Linjie Li, Michael Qizhe Shieh, Yejin Choi, Ranjay Krishna, and Yu Cheng. Thinkmorph: Emergent properties in multimodal interleaved chain-of-thought reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 141405–141447, 2026.

Yijie Guo, Dexiang Hong, Weidong Chen, Zihan She, Cheng Ye, Xiaojun Chang, and Zhendong Mao. Emoverse: A mllms-driven emotion representation dataset for interpretable visual emotion analysis. *arXiv preprint arXiv:2511.12554*, 2025.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13418–13427, 2024.

Jitesh Jain, Jianwei Yang, and Humphrey Shi. Vcoder: Versatile vision encoders for multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27992–28002, 2024.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

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 *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26763–26773, 2024.

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. In *Proceedings of the 42nd International Conference on Machine Learning*. ML Research Press, 2025a.

Zheng Lian, Haiyang Sun, Licai Sun, Haoyu Chen, Lan Chen, Hao Gu, Zhuofan Wen, Shun Chen, Zhang Siyuan, Hailiang Yao, et al. Ov-mer: Towards open-vocabulary multimodal emotion recognition. In *International Conference on Machine Learning*, pp. 37015–37050. PMLR, 2025b.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *International Conference on Learning Representations*, volume 2024, pp. 57689–57733, 2024a.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26286–26296. IEEE, 2024b.

Zuyan Liu, Yuhao Dong, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Oryx mllm: On-demand spatial-temporal understanding at arbitrary resolution. In *International Conference on Learning Representations*, volume 2025, pp. 85485–85507, 2025.

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 *European Conference on Computer Vision*, pp. 594–612. Springer, 2018.

Kuan-Chuan Peng, Tsuhan Chen, Amir Sadovnik, and Andrew C Gallagher. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 860–868, 2015.

Zheng Qin, Ruobing Zheng, Yabing Wang, Tianqi Li, Yi Yuan, Jingdong Chen, and Le Wang. Humansense: From multimodal perception to empathetic context-aware responses through reasoning mllms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 24973–24981, 2026.

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

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.

Peipei Song, Dan Guo, Xun Yang, Shengeng Tang, and Meng Wang. Emotional video captioning with vision-based emotion interpretation network. *IEEE Transactions on Image Processing*, 33: 1122–1135, 2024.

Barrett Tang, Zile Huang, Chengzhi Liu, Qiang Sun, Harry Yang, and Ser-Nam Lim. Intervening anchor token: Decoding strategy in alleviating hallucinations for mllms. In *International Conference on Learning Representations*, volume 2025, pp. 27745–27776, 2025.

Chenxi Wang, Xiang Chen, Ningyu Zhang, Bozhong Tian, Haoming Xu, Shumin Deng, and Huajun Chen. Mllm can see? dynamic correction decoding for hallucination mitigation. In *International Conference on Learning Representations*, volume 2025, pp. 13712–13736, 2025a.

Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, et al. Evaluation and analysis of hallucination in large vision-language models. *arXiv preprint arXiv:2308.15126*, 2023.

Liping Wang, Cheng Ye, Weidong Chen, Peipei Song, Bo Hu, and Zhendong Mao. A multi-agent framework with structured reasoning and reflective refinement for multimodal empathetic response generation. *arXiv preprint arXiv:2604.18988*, 2026.

Yusong Wang, Xuanye Fang, Huifeng Yin, Dongyuan Li, Guoqi Li, Qi Xu, Yi Xu, Shuai Zhong, and Mingkun Xu. Big-fusion: Brain-inspired global-local context fusion framework for multimodal emotion recognition in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 1574–1582, 2025b.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45(4):1191–1207, 2013.

Bangjun Xiao, Bingquan Xia, Bo Yang, Bofei Gao, Bowen Shen, Chen Zhang, Chenhong He, Chiheng Lou, Fuli Luo, Gang Wang, et al. Mimo-v2-flash technical report. *arXiv preprint arXiv:2601.02780*, 2026.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26596–26605, 2024.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19005–19015, 2023a.

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 *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 20383–20394, 2023b.

Qingyue Yang, Jie Wang, Xing Li, Yinqi Bai, Xialiang Tong, Huiling Zhen, Jianye Hao, Mingxuan Yuan, and Bin Li. Why attention patterns exist: A unifying temporal perspective analysis. *arXiv preprint arXiv:2601.21709*, 2026.

Ruilin Yao, Bo Zhang, Jirui Huang, Xinwei Long, Yifang Zhang, Tianyu Zou, Shili Xiong, Yi Rong, Yufei Wu, Shichao Su, et al. Lens: Multi-level evaluation of multimodal reasoning with large language models. In *International Conference on Learning Representations*, volume 2026, pp. 2627–2654, 2026a.

Zhiyuan Yao, Zheren Fu, Zhixiao Zheng, Jiajun Li, Yi Tu, and Zhendong Mao. Adapt: Attention dynamics alignment with preference tuning for faithful mllms. In *European Conference on Computer Vision*, pp. 509–526. Springer, 2026b.

Cheng Ye, Weidong Chen, Jingyu Li, Lei Zhang, and Zhendong Mao. Dual-path collaborative generation network for emotional video captioning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 496–505, 2024.

Cheng Ye, Weidong Chen, Bo Hu, Lei Zhang, Yongdong Zhang, and Zhendong Mao. Improving video summarization by exploring the coherence between corresponding captions. *IEEE Transactions on Image Processing*, 2025a.

Cheng Ye, Weidong Chen, Peipei Song, Xinyan Liu, Lei Zhang, and Zhendong Mao. Multi-round mutual emotion-cause pair extraction for emotion-attributed video captioning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 3320–3329, 2025b.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. In *International Conference on Learning Representations*, volume 2024, pp. 57153–57180, 2024.

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. *Advances in Neural Information Processing Systems*, 38:113222–113244, 2026.

Zhenlong Yuan, Xiangyan Qu, Chengxuan Qian, Rui Chen, Jing Tang, Lei Sun, Xiangxiang Chu, Dapeng Zhang, Yiwei Wang, Yujun Cai, et al. Video-star: Reinforcing open-vocabulary action recognition with tools. In *International Conference on Learning Representations*, volume 2026, pp. 51445–51468, 2026.

Junpeng Yue, Xinrun Xu, Börje F Karlsson, and Zongqing Lu. Mllm as retriever: Interactively learning multimodal retrieval for embodied agents. In *International Conference on Learning Representations*, volume 2025, pp. 31551–31580, 2025.

Fan Zhang, Zebang Cheng, Chong Deng, Haoxuan Li, Zheng Lian, Qian Chen, Huadai Liu, Wen Wang, Yi-Fan Zhang, Renrui Zhang, et al. Mme-emotion: A holistic evaluation benchmark for emotional intelligence in multimodal large language models. *arXiv preprint arXiv:2508.09210*, 2025.

Haoji Zhang, Xin Gu, Jiawen Li, Chixiang Ma, Sule Bai, Chubin Zhang, Bowen Zhang, Zhichao Zhou, Dongliang He, and Yansong Tang. Thinking with videos: Multimodal tool-augmented reinforcement learning for long video reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32903–32914, 2026.

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

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. In *International Conference on Learning Representations*, volume 2024, pp. 56969–56998, 2024.

A RELATED WORK

A.1 MULTIMODAL EMOTION REASONING

Multimodal emotion reasoning (MER) aims to infer emotional states and human intentions from multimodal information, including textual language, behavioral actions, speech signals, and social context. Initially, the community treats MER merely as a simple classification task. Some researchers develop modality fusion methods to enhance the ability to predict emotion categories (Yang et al., 2023a; Cheng et al., 2023; Wang et al., 2025b). However, given the complexity of human emotions, such simple fitting to ground truth lacks emotional interpretability, making it difficult for models to acquire genuine emotional reasoning capabilities. Recently, researchers shift the focus toward open-ended and interpretable emotional reasoning. AffectGPT (Lian et al., 2025a) constructs a descriptive emotion dataset EMER-Coarse with 2K fine-grained emotion categories and designs a two-stage training framework to better align with manually-checked results. OV-MER (Lian et al., 2025b) proposes a novel paradigm to enable emotion prediction without being confined to predefined spaces and presents a newly curated database, novel evaluation metrics, and a preliminary benchmark. EmotionLLaMA (Cheng et al., 2024) integrates multimodal inputs and aligns multimodal features with instruction tuning to enhance the emotion reasoning. Furthermore, some studies have attempted to leverage reinforcement learning algorithms to bolster emotional reasoning abilities. R1-Omni (Zhao et al., 2025) presents the first application of RL to an Omni-multimodal LLM for MER task, significantly enhancing the reasoning and generalization ability. OmniOPSD (Cheng et al., 2026) utilizes the generated rationale a s privileged evidence accessible only to the teacher model, providing dense token-level scoring and supervision for the self-generated trajectories of student model. EMO-R3 (Fang et al., 2026) proposes a reflective reinforcement learning framework, which leverages structured emotional thinking and reflective emotional reward to guide the model to perform emotion reasoning in an interpretable and step-by-step manner. Despite overcoming the limitations of closed-set prediction, these studies still treat MER as a label-level task, overlooking the subjective nature and continuous distribution of emotions. Semantically similar emotions often coexist. Reliance on discrete emotion labels prevents models from learning to reason about emotions in a continuous manner. To address this limitation, DSPO focuses on group-level soft distributional matching by aggregating all generated rollouts into a joint empirical distribution, encouraging valid emotional diversity and enhancing continuous emotional reasoning.

A.2 HALLUCINATION MITIGATION IN MLLMS

As the generative capabilities of MLLMs advance, the issue of multimodal hallucinations has become increasingly pronounced, which refers to the inconsistency between the generated text and the provided images (Ye et al., 2025a; Chen et al., 2026b). This phenomenon may stem from an over-reliance on language priors, erroneous visual perception, or inadequate cross-modal reasoning (Jain et al., 2024; Li et al., 2024; Wang et al., 2023). Researchers have explored various strategies to mitigate these hallucinations. Fine-tuning approaches focus on constructing high-quality datasets for fine-grained alignment to bridge the gap between visual and textual knowledge (You et al., 2024; Liu et al., 2024a). However, this demands valuable annotation costs and substantial computational resources. Alternatively, post-hoc methods utilize external tools or self-reflection mechanisms to correct hallucinated outputs (Zhou et al., 2024; Huang et al., 2024). Moreover, certain decoding strategies delve into detecting anomalous attention tokens during generation, applying targeted interventions based on these observed patterns (Tang et al., 2025; Wang et al., 2025a).

Crucially, for multimodal emotion reasoning tasks, the exacerbation of multimodal hallucinations is remarkably severe. We infer this is because emotional cues are implicitly nested within abstract semantics, such as subtle micro-expressions, lighting, or overall atmospheric nuances, rather than explicit physical entities (Ye et al., 2024; Song et al., 2024; Ye et al., 2025b). Without specifically fine-tuning, existing MLLMs lack the intrinsic capability to mine these implicit emotion cues, leading to the frequent fabrication of visual facts to cater to emotional conclusions. To overcome this critical bottleneck, we introduce a Counterfactual Visual Intervention Gating (CVIG). By masking specific visual regions, CVIG generates a counterfactual image and computes the causal discrepancy in emotion prediction probabilities to evaluate the causal impact of the proposed visual cues. By rewarding genuinely causal visual evidence and penalizing hallucinated fabrications, CVIG effectively mitigates the multimodal hallucinations in emotion reasoning.

A.3 Proof: DSPO is a Variational Lower Bound of Genuine Human Emotion

In this section, we make a theoretical analysis (Shannon, 1948) to prove that our proposed DSPO is a variational lower bound of genuine human emotion from an information-theoretic perspective. We follow the notations above: the image input $\mathcal{V}$, the output of MLLM $\mathcal{O}$, and the ground-truth human emotional distribution $\mathcal{Z}$, respectively. Overall, regarding our optimization objective:

$$\mathcal{J}(\theta) = \mathbb{E}_{(\mathcal{V},Q)} E_{\mathcal{O}\sim\pi_{\theta_{old}}} \left[\frac{1}{G} \sum_{i=1}^{G} \Big(\min \Big[\rho_i(\theta)\hat{A}_i,\ \text{clip}\left(\rho_i(\theta), 1-\epsilon, 1+\epsilon\right) \hat{A}_i \Big] \right] - \beta \mathbb{D}_{KL}(\pi_\theta \,||\, \pi_{ref}) \tag{1}$$

we aim for the text generated by the MLLM to exhibit the highest similarity with the true human emotion distribution given the visual prior, which is equivalent to maximizing the conditional mutual information $I(\mathcal{O}; \mathcal{Z}|\mathcal{V})$. Mathematically, it can be decomposed into the following form:

$$I(\mathcal{O}; \mathcal{Z}|\mathcal{V}) = H(\mathcal{Z}|\mathcal{V}) - H(\mathcal{Z}|\mathcal{O}, \mathcal{V}), \tag{2}$$

$H(\mathcal{Z}|\mathcal{V})$ denotes the inherent uncertainty of human emotion given the image cues, which is a constant determined by human priors. Thus, our goal is to minimize $H(\mathcal{Z}|\mathcal{O}, \mathcal{V})$, which represents the residual uncertainty of human emotion, given the provided image cues and the response of MLLMs. Based on the information-theoretic definition, it is equivalent to maximizing the following expectation:

$$-H(\mathcal{Z}|\mathcal{O}, \mathcal{V}) = \mathbb{E}_{\mathcal{O},\mathcal{Z}}[\log P_{true}(\mathcal{Z}|\mathcal{O}, \mathcal{V})], \tag{3}$$

for $P_{true}(\mathcal{Z}|\mathcal{O}, \mathcal{V})$, it is an internal representation that is difficult to observe directly. Thus, we introduce Evidence Lower Bound (ELBO) (Kingma & Welling, 2013) to approximate it.

$$\begin{aligned} \mathbb{E}_{\mathcal{O},\mathcal{Z}}[\log P_{true}(\mathcal{Z}|\mathcal{O}, \mathcal{V})] \geq & \underbrace{\mathbb{E}_{\mathcal{O}\sim\pi_\theta}\left[\mathbb{E}_{\mathcal{Z}\sim P^*(\mathcal{Z}|\mathcal{V})}[\log Q(\mathcal{Z}|\mathcal{O})]\right]}_{\text{Reconstruction Term}} \\ & - \underbrace{\mathbb{E}_{\mathcal{O}\sim\pi_\theta}\left[D_{KL}(P_{true}(\mathcal{Z}|\mathcal{V}), ||, P_{prior}(\mathcal{Z}))\right]}_{\text{Prior Penalty Term}}, \end{aligned} \tag{4}$$

since both $P_{true}(\mathcal{Z}|\mathcal{V})$ and $P_{prior}(\mathcal{Z})$ are human prior distributions unrelated to π_θ, the KL-divergence term is a non-negative constant. We focus on maximizing the reconstruction term:

$$I(\mathcal{O}; \mathcal{Z}|\mathcal{V}) \geq \underbrace{\mathbb{E}_{\mathcal{O}\sim\pi_\theta}}_{\text{CVIR}} \underbrace{\left[\mathbb{E}_{\mathcal{Z}\sim P^*(\mathcal{Z}|\mathcal{V})}[\log Q(\mathcal{Z}|\mathcal{O})]\right]}_{\text{DEDR}} \sim \mathcal{J}_{\text{DSPO}}(\theta), \tag{5}$$

for the reconstruction term, $\mathbb{E}_{\mathcal{O}\sim\pi_\theta}$ denotes the ability for the MLLM to generate factually accurate descriptions, which refers to CVIR. Besides, $\mathbb{E}_{\mathcal{Z}\sim P^*(\mathcal{Z}|\mathcal{V})}[\log Q(\mathcal{Z}|\mathcal{O})]$ denotes the ability to use MLLM outputs to fit the true human distribution, which refers to DEDR. Thus, our proposed DSPO is fundamentally a variational lower bound of genuine human emotion.

A.4 Experimental Setup

Datasets and Metrics. We evaluate the emotion reasoning of MLLMs on three public benchmarks, *i.e.,* EmoSet (Yang et al., 2023b), Emotion6 (Peng et al., 2015), and WebEmo (Panda et al., 2018). We evaluate DSPO on both in-domain and out-of-domain (OOD) benchmarks. Specifically, we leverage EmoSet/Emotion6 as the training source and the other two datasets as the external datasets. For each dataset, we use emotion accuracy as the evaluation metric. All reported accuracy metrics are computed via hard matching of the final \boxed{} prediction against the ground-truth label.

Base Model and Implementation Details. We compare DSPO by two backbones: LLaVA-1.5-7B (Liu et al., 2024b) and Qwen2.5-VL-3B-Instruct (Bai et al., 2025), and with zero-shot inference, SFT, two reinforcement-learning baselines *i.e.,* GRPO (Shao et al., 2024) and DAPO (Yu et al., 2026), and a SOTA method EMO-R3 (Fang et al., 2026) under two rollout budgets. We employ two evaluation settings: 1) EmoSet (Yang et al., 2023b) for in-domain evaluation and Emotion6 (Peng et al., 2015)/WebEmo (Panda et al., 2018) for out-of-domain evaluation and 2) Emotion6 for in-domain evaluation and EmoSet/WebEmo for out-of-domain evaluation. Following the standard hyper-parameter configurations established in prior GRPO-based works Shao et al. (2024); Fang et al. (2026), we set the clipping parameter to $\epsilon = 0.2$ and the KL penalty coefficient to $\beta = 0.01$. For the ground-truth subjective distribution construction, we directly use the original multi-annotator

emotion probability distributions for the Emotion6 dataset. For EmoSet and WebEmo, we leverage the Warriner & NRC VAD lexicon (∼54,800 lemmas) for static priors, and Mimo-v2 Xiao et al. (2026) for dynamic context captioning. The number of rollouts per group is set to $G = 4/8$, and the number of emotion candidates per rollout is $N = 5$ unless otherwise specified. The default reward coefficient is $\lambda_{acc} = 1.0, \lambda_{fmt} = 0.1, \lambda_{cf} = 0.1$, and $\lambda_{div} = 0.3$. All experiments are conducted on $8\times$ A800 (80GB) GPUs.

Distribution-based Evaluation. To intuitively quantify whether DSPO truly learns the authentic emotion distribution, we evaluate on four evaluation metrics based on emotion distribution. First, we use EmoSet as the training source and evaluate the models on the multi-annotator distribution labels of Emotion6. Furthermore, since DSPO and previous methods both generate only a single final emotion category per inference, we perform $M = 50$ independent samplings for each image to approximate the distribution:

$$\text{Sample}_M(i) = \{y_i^1, y_i^2, \ldots, y_i^M\}, \quad q_i(c) = \frac{m_{i,c} + \alpha}{M + \alpha|C|}, \tag{6}$$

where y_i^t denotes the emotion category predicted for the image i in the t-th sampling. $m_{i,c}$ is the number of times category c appears in M inferences, where $c \in C = \{Anger, Disgust, Fear, Joy, Sadness, Surprise\}$ is one of the six emotion labels in Emotion6. α is a smoothing coefficient introduced to prevent divergence collapse caused by emotional probabilities of zero. Besides, for the ground-truth emotion distribution, we suppose that for image i there are N_i multi-label annotations, and category c receives $n_{i,c}$ annotations. The ground-truth emotion distribution could be expressed as $p_i(c) = \frac{n_{i,c}}{N_i}$.

Distribution-based Metrics. First, we consider using divergence-based metrics to measure the similarity between the predicted distribution and the ground-truth distribution. Specifically, we employ both KL and JS divergence due to the instability of KL divergence:

$$D_{KL}(p_i||q_i) = \sum_{c=1}^{|C|} p_i(c) log \frac{p_i(c)}{q_i(c)}, \tag{7}$$

$$D_{JS}(p_i.q_i) = \frac{D_{KL}(p_i||m_i) + D_{KL}(q_i||m_i)}{2}, \quad m_i = \frac{p_i + q_i}{2}, \tag{8}$$

besides, we also employ two entropy-based metrics to evaluate whether the uncertainty of the predicted distribution approximates the ground-truth:

$$H_i^{human} = -\frac{1}{log|C|} \sum_{c=1}^{|C|} p_i(c) log p_i(c), \quad H_i^{model} = -\frac{1}{log|C|} \sum_{c=1}^{|C|} q_i(c) log q_i(c), \tag{9}$$

$$\text{En-MAE} = |H_i^{model} - H_i^{human}|, \tag{10}$$

we will report the absolute entropy values of the models H_i^{model} and their proximity to the ground-truth entropy En-MAE.

A.5 Causal-grounded Emotional Thinking Template

Causal-grounded Emotional Thinking:

<step1>*Identify the core visual element (action, facial expression, object, or environment) that triggers the emotion. Finally, provide a bounding box coordinate in the format* $[y_{min}, x_{min}, y_{max}, x_{max}]$ *put in* \bboxed{}.</step1>
<step2>*Reflect on the psychological state. Describe in detail how a human observer would emotionally resonate with this trigger, specifically expressing the valence and arousal of the feeling.*</step2>
<step3>*Synthesize the visual evidence and psychological reflection to generate K distinct emotions. For each, provide an emotion label and a brief justification linking back*

to the visual trigger. Finally, provide an emotion list in the format $\{e_1, e_2, \ldots, e_K\}$ *put in* `\list{}.</step3>`
`<step4>`*Based on the above analysis, choose the most appropriate option from the following emotional descriptions:*

[Emotion Set of Dataset]

The chosen option MUST BE put in `\boxed{}.</step4>`