ENHANCING MULTI-MODAL LLMS REASONING VIA DIFFICULTY-AWARE GROUP NORMALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement Learning with Verifiable Rewards (RLVR) and Group Relative Policy Optimization (GRPO) have significantly advanced the reasoning capabilities of large language models. Extending these methods to multimodal settings, however, faces a critical challenge: the instability of *std*-based normalization, which is easily distorted by extreme samples with nearly positive or negative rewards. Unlike pure-text LLMs, multimodal models are particularly sensitive to such distortions, as both perceptual and reasoning errors influence their responses. To address this, we characterize each sample by its **difficulty**, defined through perceptual complexity (measured via visual entropy) and reasoning uncertainty (captured by model confidence). Building on this characterization, we propose **difficulty-aware group normalization**, which re-groups samples by difficulty levels and shares the *std* within each group. Our approach preserves GRPO's intra-group distinctions while eliminating sensitivity to extreme cases, yielding significant performance gains across multiple multimodal reasoning benchmarks.

1 INTRODUCTION

Reinforcement Learning with Verifiable Rewards (RLVR) has enabled significant advances in the reasoning capabilities of both large language models (LLMs) (DeepSeek-AI et al., 2025; Yang et al., 2025a; Lambert et al., 2024) and multi-modal large language models (MLLMs) (Zhang et al., 2025b; Huang et al., 2025a). Within this paradigm, Group Relative Policy Optimization (GRPO) (Shao et al., 2024b) demonstrates strong performance by applying standard deviation (*std*)-based normalization to rewards within each response group. This *std*-based normalization rescales intra-group distinctions between positive and negative responses, thereby stabilizing training.

Despite these advances, the *std*-based normalization suffers from a critical limitation: *sensitive to extreme samples* — those with response groups that are almost entirely positive or negative. Specifically, when rewards in a group collapse to near 0 or 1, the resulting low *std* overemphasizes the extreme samples during optimization. Meanwhile, samples with more balanced rewards are neglected, leading to imbalanced optimization. This issue is particularly pronounced in MLLMs, where the complexity of multimodal inputs increases the occurrence of such extreme samples. As illustrated in Figure 1, MLLM responses are jointly influenced by challenges from perceptual complexity and reasoning uncertainty, making them more susceptible to extreme reward distributions.

While removing the *std* term mitigates the risk of overfitting to extreme samples (Liu et al., 2025b), it simultaneously discards the valuable intra-group distinctions, which are essential for effective and stable optimization. Therefore, *the key issue lies not in the std-normalization itself, but rather in the way groups are constructed*: when group size is small, extreme cases become inevitable. Enlarging group sizes during rollouts could help, but it incurs prohibitive computational costs.

Motivated by this, we propose to account for the challenges of various samples, which we refer to as **difficulty-aware re-grouping**. We characterize each sample's difficulty from two complementary perspectives: (i) a data-centric view, where high entropy in the image reflects *perceptual difficulty*; and (ii) a model-centric view, where low confidence in model responses reflects *reasoning difficulty*. By re-grouping samples according to these difficulty levels and sharing the *std* within each group, our method preserves intra-group distinctions while mitigating sensitivity to extreme cases. Specifically, our difficulty-based re-group strategy is achieved by:

Perceptual difficulty based regrouping. We quantify perceptual difficulty through spectral analysis of image patch covariances, where higher entropy in the resulting eigenvalue distribution indicates greater visual complexity. Intuitively, images with more diverse and complex visual patterns exhibit higher entropy, reflecting greater perceptual difficulty.

Reasoning difficulty based regrouping. Leveraging the insight that token-level log probabilities reflect reasoning confidence (Yu et al., 2025c), we measure reasoning difficulty through the model's token-level confidence, where lower average log probabilities indicate greater uncertainty in generating correct reasoning chains, reflecting higher reasoning difficulty.

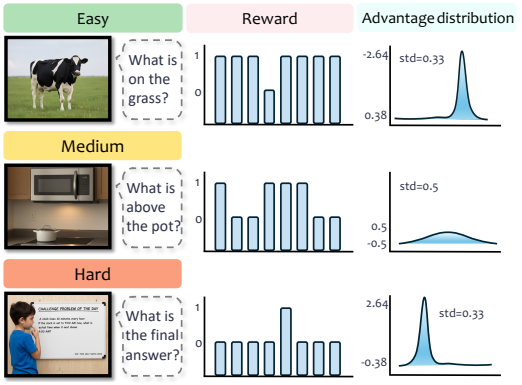


Figure 1: Advantage distribution after the normalization of reward varies among samples. Extreme samples like easy and hard ones are amplified after *std*-normalization, whereas medium samples exhibit more balanced advantages.

By explicitly decomposing difficulty into data-centric (**perceptual**) and model-centric (**reasoning**) groups, our method allows each group of samples to share separate *std*s for perceptual and reasoning aspects. These normalized advantages are then combined via element-wise multiplication, effectively integrating intrinsic data complexity and model uncertainty, ensuring stable optimization that preserves meaningful intra-group distinctions. Our contributions are summarized into the following three aspects:

- We identify that the limitation of *std*-normalization in group-based RL methods (*e.g.* GRPO and DAPO) originates from group construction, particularly in multimodal reasoning tasks, and propose a difficulty-based re-grouping strategy to build more robust groups.
- We explicitly decompose difficulty into perceptual and reasoning aspects, and integrate them via element-wise combination, effectively capturing both data complexity and model uncertainty while preserving intra-group distinctions.
- Our proposed strategy achieves significant performance gains. Concretely, building upon GRPO and DAPO, our strategy attains more than 2% average performance improvements.

2 Preliminary

In this section, we briefly introduce the key concepts and training setup for multimodal reasoning under RLVR (DeepSeek-AI et al., 2025). We first formulate the task, and then revisit the standard GRPO framework (Shao et al., 2024b) along with its improved variant, Decoupled Clip and Dynamic Sampling Policy Optimization (DAPO) (Yu et al., 2025b).

2.1 Task Formulation

We consider the problem of multimodal reasoning under the RLVR paradigm. Let $\{\mathcal{I}, \mathcal{Q}\} \in \mathcal{D}$ denote a multimodal input, where the dataset $\mathcal{D}$ includes image $\mathcal{I}$ and text question $\mathcal{Q}$. The model generates a reasoning response o given $\{\mathcal{I}, \mathcal{Q}\}$ and receives a verifiable reward r based on the correct answer y (Wu et al., 2025; Wang et al., 2025a). The response o typically contains both the reasoning steps and the final answer, with the reasoning steps enclosed in `<think>...</think>` and the final answer enclosed in `\boxed{}`. We employ a binary reward function, where $r(o, y) = 1$ if the final answer is equal to the correct answer y, and $r(o, y) = 0$ otherwise. The reasoning process is modeled as a policy $\pi_\theta(o|\mathcal{I}, \mathcal{Q})$ parameterized by θ to maximize the expected reward:

$$\mathcal{J}_{\text{RLVR}}(\theta) = \max_\theta \mathbb{E}_{\{\mathcal{I},\mathcal{Q}\}\sim\mathcal{D}} \mathbb{E}_{o\sim\pi_\theta(\cdot|\mathcal{I},\mathcal{Q})}[r(o, y)]. \tag{1}$$

Our goal is to enhance the reasoning capabilities of an instruction-tuned MLLM, thereby significantly improving its performance on downstream multimodal reasoning tasks.

2.2 Core Algorithms of Reinforcement Learning with Verifiable Reward

Group Relative Policy Optimization (GRPO) is derived from Proximal Policy Optimization (PPO) (Schulman et al., 2017), with the key distinction that GRPO replaces the advantage estimates obtained via Generalized Advantage Estimation (GAE) with group-relative advantages computed from a group of outputs.

Specifically, for each input $\mathcal{I}$, $\mathcal{Q}$, GRPO samples a group of outputs $\{o_1, o_2, \ldots, o_G\}$ from the old policy model $\pi_{\theta_{\text{old}}}$, with rollout size G. The advantage of the i-th response is computed by normalizing the rewards among the group:

$$\hat{A}_i = \frac{r_i - mean(\{r_1, r_2, \ldots, r_G\})}{std(\{r_1, r_2, \ldots, r_G\})}. \tag{2}$$

GRPO adopts a clipped objective, together with a directly imposed KL penalty term:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{(\mathcal{I},\mathcal{Q})\sim\mathcal{D},\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(o|\mathcal{I},\mathcal{Q})} \left\{ \frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|} \right. \tag{3}$$
$$\left. \min\left[\frac{\pi_\theta(o_{i,t} \mid \mathcal{I},\mathcal{Q},o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid \mathcal{I},\mathcal{Q},o_{i,<t})}\hat{A}_{i,t},\ \text{clip}\Big(\frac{\pi_\theta(o_{i,t} \mid \mathcal{I},\mathcal{Q},o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid \mathcal{I},\mathcal{Q},o_{i,<t})}, 1-\epsilon, 1+\epsilon\Big)\hat{A}_{i,t}\right] - \beta\mathbb{D}_{KL}(\pi_\theta\|\pi_{\text{ref}})\right\}.$$

ϵ is the hyperparameter to control the clipping range of the importance sampling ratio, and β is the penalty strength of how far the current policy π_θ deviates from the reference policy π_{ref}.

Decoupled Clip and Dynamic Sampling Policy Optimization (DAPO) is a variant of GRPO adopting an asymmetric clipping range with a larger upper bound, dynamic sampling, token-level policy gradient loss, and overlong reward shaping. The objective function of DAPO is defined as:

$$\mathcal{J}_{\text{DAPO}}(\theta) = \mathbb{E}_{(\mathcal{I},\mathcal{Q})\sim\mathcal{D},\,\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(o|\mathcal{I},\mathcal{Q})} \left\{ \frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|} \right. \tag{4}$$
$$\left. \min\left[\frac{\pi_\theta(o_{i,t} \mid \mathcal{I},\mathcal{Q},o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid \mathcal{I},\mathcal{Q},o_{i,<t})}\hat{A}_{i,t},\ \text{clip}\Big(\frac{\pi_\theta(o_{i,t} \mid \mathcal{I},\mathcal{Q},o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid \mathcal{I},\mathcal{Q},o_{i,<t})}, 1-\epsilon_{\text{low}}, 1+\epsilon_{\text{high}}\Big)\hat{A}_{i,t}\right]\right\}.$$

3 Method

In this section, we introduce our difficulty-based regroup strategy in detail. We first represent our perceptual difficulty-based regrouping in Section 3.1, then we describe our reasoning difficulty-based regrouping in Section 3.2. Finally, we show our combination strategies.

3.1 Perceptual difficulty-based re-grpouping

Perceptual difficulty estimation. To estimate the perceptual difficulty of a batch $\mathcal{B} = \{(\mathcal{I}_s, \mathcal{Q}_s)\}_{s=1}^{B}$, we first extract patch-level visual features from the Qwen2.5-VL-7B visual encoder $\mathbf{\Omega}_v$:

$$\mathbf{F}_s = \mathbf{\Omega}_v(\mathcal{I}_s) \in \mathbb{R}^{P\times d} = [\boldsymbol{f}_s^1, \boldsymbol{f}_s^2, \cdots, \boldsymbol{f}_s^P]^\top, \tag{5}$$

where P denotes the number of spatial patches and d is the feature dimension, with $\boldsymbol{f}_s^j \in \mathbb{R}^{d\times 1}, j = 1, \cdots, P$ representing the feature of the j-th patch.

Compared to CLIP-based representations (Radford et al., 2021), these patch-level features not only capture finer spatial granularity that preserves local details, but also align better with the downstream textual decoder $\mathbf{\Omega}_t$, ensuring both stability and semantic consistency.

We then compute the empirical covariance matrix to capture intra- and inter-patch variances:

$$\mathbf{C}_s = \tfrac{1}{P-1}(\mathbf{F}_s - \mathbf{1}_P\boldsymbol{\mu}_s^\top)(\mathbf{F}_s - \mathbf{1}_P\boldsymbol{\mu}_s^\top)^\top, \quad \boldsymbol{\mu}_s = \tfrac{1}{P}\sum_{j=1}^{P}\boldsymbol{f}_s^j. \tag{6}$$

where $\mathbf{1}_P$ is a $P\times 1$ column vector of ones and $\boldsymbol{\mu}_s$ is the mean of the patches feature of the $\mathcal{I}_s$. The diagonal entries measure the variance of each feature dimension across patches, while the off-diagonal

terms capture correlations between different feature dimensions. This covariance structure reveals whether visual features are dominated by a few strong dimensions or by multiple interacting factors, providing a principled basis for assessing perceptual difficulty.

Since $\mathbf{C}_s$ is a symmetric positive semidefinite matrix, we then perform eigenvalue decomposition for spectral analysis:

$$\mathbf{C}_s = \mathbf{V}_s \mathbf{\Lambda}_s \mathbf{V}_s^\top, \quad \mathbf{\Lambda}_s = \mathrm{diag}(\lambda_s^1, \ldots, \lambda_s^P),\ \lambda_s^k \geq 0. \tag{7}$$

λ_s^k denotes the k-th eigenvalue, quantifying the variance along one orthogonal principal direction. Concentrated eigenvalues indicate that most variance is captured by a few dimensions, whereas more balanced eigenvalues imply richer visual structure and higher visual complexity.

The final perceptual difficulty score is defined as the entropy (Shannon, 1948) of the normalized distribution of eigenvalues:

$$H(\mathcal{I}_s) = -\sum_{k=1}^{P} p_s^k \log p_s^k, \tag{8}$$

where p_s is the normalized probability distribution of the eigenvalues, and each element in p_s is calculated as:

$$p_s^k = \frac{\lambda_s^k}{\sum_{j=1}^{P} \lambda_s^j}, \quad \text{with} \sum_{k=1}^{P} p_s^k = 1. \tag{9}$$

Here, low entropy corresponds to the visually easy sample, with variance concentrated on a few dominant components, whereas high entropy indicates the visually difficult sample, with variance distributed across many patches.

Perceptual difficulty-based regrouping. Given the perceptual difficulty scores within a batch, we partition samples into three groups using the 25th and 75th percentiles $\tau_{0.25}$ and $\tau_{0.75}$:

$$\mathcal{S}_1 = \{s \mid H(\mathcal{I}_s) \leq \tau_{0.25}\}, \quad \mathcal{S}_2 = \{s \mid \tau_{0.25} < H(\mathcal{I}_s) < \tau_{0.75}\}, \quad \mathcal{S}_3 = \{s \mid H(\mathcal{I}_s) \geq \tau_{0.75}\}. \tag{10}$$

For each group a, the reward set can be defined as:

$$\mathcal{R}_a = \{r_{s,i} \mid i = 1, \cdots, G, s \in \mathcal{S}_a\},\ a \in \{1, 2, 3\} \tag{11}$$

where $r_{s,i}$ refers to the i-th reward of the s-th sample which belongs to the group $\mathcal{S}_a$.

We then compute the shared standard deviation $std(\mathcal{R}_a)$ of group rewards, and normalize the reward of each sample in batch with the new $std(\mathcal{R}_a)$ to calculate advantage accordingly:

$$A_{s,i}^{\text{Perceptual}} = \frac{r_{s,i} - mean(r_{s,1}, r_{s,2}, \ldots r_{s,G})}{std(\mathcal{R}_a)}, \tag{12}$$

$$std(\mathcal{R}_a) = \sqrt{\frac{1}{|\mathcal{R}_a| - 1} \sum_{r_{s,i} \in \mathcal{R}_a} \left(r_{s,i} - \frac{1}{|\mathcal{R}_a|} \sum_{r_{s,i} \in \mathcal{R}_a} r_{s,i}\right)^2}, \tag{13}$$

where $|\mathcal{R}_a|$ denotes to the cardinality of $\mathcal{R}_a$.

By grouping samples into low-, medium-, and high-entropy categories, the normalization scale is shared only among samples with comparable perceptual difficulty. This mitigates the influence of extreme samples, balances treatment across different levels of visual complexity, and ultimately stabilizes optimization.

3.2 Reasoning difficulty-based regrouping

Reasoning difficulty estimation. While perceptual difficulty captures the intrinsic complexity of the image, reasoning difficulty is shaped by the model's intrinsic confidence in generating the final answer. Even for inputs with similar visual complexity, the model may exhibit varying confidence levels: high confidence (assigning a high probability to reasoning chains) implies a clear and reliable reasoning path, whereas low confidence indicates uncertainty and potential reasoning failures. Following this intuition, we quantify reasoning difficulty using the model's probabilities to its reasoning chains.

For the given batch $\mathcal{B} = \{(\mathcal{I}_s, \mathcal{Q}_s)\}_{s=1}^{B}$, and the generated G responses for each sample, we denote the i-th response as $o_{s,i} = (o_{s,i}^1, \ldots, o_{s,i}^T)$, where $o_{s,i}^n$ is the n-th token and T is the sequence length.

Based on the token-level log probability $\pi_\theta\left(o_{s,i}^n \,\middle|\, \mathcal{I}_s, \mathcal{Q}_s, o_{s,i}^{<n}\right)$, we aggregate across tokens to obtain the sequence-level log probability for response $o_{s,i}$:

$$L_{s,i} = \sum_{n=1}^{T} \pi_\theta\left(o_{s,i}^n \,\middle|\, \mathcal{I}_s, \mathcal{Q}_s, o_{s,i}^{<n}\right). \tag{14}$$

Then we define the model confidence for sample $(\mathcal{I}_s, \mathcal{Q}_s)$ as the average sequence-level log probability across its G rollouts:

$$L(\mathcal{Q}_s) \;=\; \frac{1}{G}\sum_{i=1}^{G} L_{s,i}. \tag{15}$$

This formulation reflects the model's internal confidence: High and consistent $L(\mathcal{Q}_s)$ indicates a reliable reasoning chain, whereas low or fluctuating $L(\mathcal{Q}_s)$ reflects epistemic uncertainty, implying that more challenging reasoning sample.

Reasoning difficulty-based regrouping. Given the model confidence scores $L(\mathcal{Q}_s)$ for the batch $\mathcal{B} = \{(\mathcal{I}_s, \mathcal{Q}_s)\}_{s=1}^{B}$, we divide samples into b groups according to the quantiles of their confidence distribution. Let $\{\tau_0, \tau_1, \ldots, \tau_b\}$ denote the quantile boundaries, with $\tau_0 = 0$ and $\tau_b = 1$. Each question $\mathcal{Q}_s$ is then assigned to a group by:

$$\mathcal{M}_u = \{\, s \mid \tau_{u-1} \le L(\mathcal{Q}_s) < \tau_u \,\}, \quad u \in \{1, \ldots, b\}. \tag{16}$$

Within each group $\mathcal{M}_u$, we define the reward set as:

$$\mathcal{R}_u = \{r_{s,i} \mid i = 1, \ldots, G, s \in \mathcal{M}_u\},\ u \in \{1, \ldots, b\} \tag{17}$$

where $r_{u,i}$ is the reward of the i-th response for the sample, which belongs to the u group. We can then calculate the shared standard deviation $std(\mathcal{R}_\mathcal{W})$ of reasoning difficulty-based group, and compute the advantage accordingly:

$$A_{s,i}^{\text{Reasoning}} = \frac{r_{s,i} - mean(r_{s,1}, r_{s,2}, \ldots r_{s,G})}{std(\mathcal{R}_u)}, \tag{18}$$

where the *std* can be calculated as:

$$std(\mathcal{R}_u) = \sqrt{\frac{1}{|\mathcal{R}_u| - 1} \sum_{r_{s,i} \in \mathcal{R}_u} \left(r_{s,i} - \frac{1}{|\mathcal{R}_u|} \sum_{r_{s,i} \in \mathcal{R}_u} r_{s,i}\right)^2}, \tag{19}$$

This regrouping ensures that responses with similar confidence levels are normalized on comparable scales, mitigating instability introduced by overconfident or underconfident samples.

3.3 Combination for Robust Optimization

To leverage the complementary aspects of perceptual and reasoning difficulty, we propose an element-wise combination strategy. Specifically, given the perceptual-based group normalized advantage $A^{\text{Perceptual}}$, the reasoning-based group normalized advantage $A^{\text{Reasoning}}$, and the original GRPO advantage A^{GRPO}, the combined advantage is defined as:

$$A^{\text{Combined}} \;=\; \alpha_{\text{Ori}} \cdot A^{\text{GRPO}} \;+\; \alpha_{\text{Percep}} \cdot A^{\text{Perceptual}} \;+\; \alpha_{\text{Reason}} \cdot A^{\text{Reasoning}}, \tag{20}$$

where $\alpha_{\text{Ori}}, \alpha_{\text{Percep}}, \alpha_{\text{Reason}}$ are weighting coefficients that balance the contributions of the three components. Perceptual difficulty, quantified by the entropy in the image, captures the *visual complexity* of multimodal inputs; reasoning difficulty, derived from token- and sequence-level log probabilities, reflects the *model uncertainty* during reasoning. Integrating these difficulty-based advantages with the original GRPO advantage allows the model to preserve meaningful intra-sample distinctions and incorporate both intrinsic and extrinsic difficulty context, providing a more stable and informative advantage for policy optimization.

Table 1: The Acc performance of two re-grouping strategies in our method. We take DAPO as our backbone. The best results are indicated in **boldface**, and the second-best results are underlined.

Model	MathVerse	MathVision	MathVista	WeMath	HallusionBench	Avg
DAPO	50.4	27.6	70.7	69.4	68.6	57.3
DAPO + Perceptual regrouping	**52.3**	28.0	71.4	70.9	70.9	58.7
DAPO + Reasoning regrouping	**52.3**	28.3	71.6	70.8	68.9	58.4
DAPO + Both	51.9	**29.0**	**72.2**	**71.8**	**71.4**	**59.3**

4 EXPERIMENT

In this section, we demonstrate the effectiveness of our re-grouping strategy. Specifically, we first show the experimental settings, including the dataset, benchmark, baselines, and implementation details. Then, we illustrate the ablation studies to analyze the two regroup strategies. Finally, we compare with the state-of-the-art methods over various benchmarks.

4.1 EXPERIMENTAL SETTINGS

Dataset. For training, we rely on the Geometry3K (Lu et al., 2021) dataset, which provides 2.1K training samples and 0.3K validation samples.

Benchmark. We evaluate our method on five benchmarks: four visual reasoning datasets, namely MathVerse (Zhang et al., 2024), MathVision (Wang et al., 2024), MathVista (Lu et al., 2024), and WeMath (Qiao et al., 2025), as well as one visual perception benchmark, HallusionBench (Guan et al., 2024). In addition, we assess the in-domain performance by comparing our method with the vanilla GRPO and DAPO.

Baseline. To evaluate the performance of our method, we consider three categories of baselines: (1) Closed-source models, such as GPT-4o (Hurst et al., 2024), and Cloud-3.5-sonnet (Anthropic, 2024). (2) Open source models: including InternVL-2.5-8B-Instruct (Chen et al., 2024), LLaVA-OneVision-7B (Li et al., 2024), Kimi-VL-16B (Du et al., 2025a), URSA-8B (Luo et al., 2025), and Mulberry-7B (Yao et al., 2024). (3) RLVR-based Models: MLLMs trained with reinforcement learning using verifiable rewards, representing the current mainstream approaches in this line of research. This category includes R1-VL-7B (Zhang et al., 2025a) , Vision-R1-7B (Huang et al., 2025b) , R1-OneVision-7B (Yang et al., 2025b) , OpenVLThinker-7B (Deng et al., 2025) , MM-Eureka-Qwen-7B (Meng et al., 2025) , ADORA-7B (Gui & Ren, 2025) , ThinkLite-7B-VL (Wang et al., 2025d) , and VLAA-Thinker-7B (Chen et al., 2025a) .

Implementation details. Following prior work (Liu et al., 2025a), we use Qwen2.5-VL-7B-Instruct (Bai et al., 2025) as our base model and adpot EasyR1 (Zheng et al., 2025) as our reinforcement learning framework. All experiments are conducted on 8 NVIDIA H20 96G GPUs. We adopt the default settings from EasyR1, using a learning rate of $1e^{-6}$, a global batch size of 128, a rollout batch size of 512, and a rollout size of 8.

4.2 ABLATION STUDIES

To better understand the contribution of each component in our method, we conduct ablation studies on five benchmarks by comparing four settings: vanilla DAPO, DAPO with perceptual regrouping, DAPO with reasoning regrouping, and our final method. Results are summarized in Table 1.

The effects of Perceptual difficulty-based regrouping. Using perceptual difficulty-based regrouping alone yields consistent performance gains across benchmarks. For instance, on HallusionBench, which is explicitly designed to evaluate perceptual ability, we observe an improvement of 3.4% over vanilla DAPO. This demonstrates that regrouping samples via spectral analysis of image patch covariances enhances the model's perceptual grounding by mitigating the dominance of extremely easy or hard cases.

The effects of Reasoning difficulty-based regrouping. When regrouping with model confidence, the average accuracy increases to 58.4, and it's notable to observe a 3.8% gain on MathVerse, even surpassing the performance of our method. This indicates that the model's internal confidence estimation also serves as a reliable signal for stabilizing optimization.

Table 2: Performance comparison of Multi-modal LLMs with over 5 benchmarks. Accuracy scores (%) are reported for all benchmarks for clarity. Data sizes used for SFT and RL are annotated in blue and red, respectively. The best value in each column is shown in **bold**, and the second-best is underlined.

Model	Data Size	MathVerse	MathVision	MathVista	WeMath	HallusionBench	Average
Close-source models							
GPT-4o	-	50.8	30.4	63.8	69.0	71.4	-
Claude-3.5-Sonnet	-	26.5	38.0	67.7	-	71.6	
Open-source models							
InternVL-2.5-8B-Instruct (Chen et al., 2024)	-	39.5	19.7	64.4	-	67.3	-
LLaVA-OneVision-7B (Li et al., 2024)	-	26.2	-	63.2	-	48.4	-
Kimi-VL-16B (Du et al., 2025a)	-	44.9	21.4	68.7	-	66.2	-
URSA-8B (Luo et al., 2025)	-	45.7	26.2	59.8	-	-	-
Mulberry-7B (Yao et al., 2024)	-	-	-	63.1	-	-	-
reinforcement learning with verifiable reward based							
R1-VL-7B (Zhang et al., 2025a)	260K+10K	40.0	24.7	63.5	-	-	-
Vision-R1-7B (Huang et al., 2025b)	200K+10K	<u>52.4</u>	-	**73.5**	-	-	-
R1-OneVision-7B (Yang et al., 2025b)	155K+10K	46.1	22.5	63.9	62.1	65.6	52.0
OpenVLThinker-7B (Deng et al., 2025)	35K+15K	48.0	25.0	71.5	67.8	70.8	56.6
MM-Eureka-Qwen-7B (Meng et al., 2025)	15K	50.5	28.3	71.5	65.5	68.3	56.8
ADORA-7B (Gui & Ren, 2025)	2.1K	50.1	27.6	71.1	67.1	53.1	53.8
ThinkLite-7B-VL (Wang et al., 2025d)	11K	50.2	27.6	72.7	69.2	71.0	58.1
VLAA-Thinker-7B (Chen et al., 2025a)	25K	49.9	26.9	68.8	67.9	68.6	56.4
NoisyRollout (Liu et al., 2025a)	2.1K	**53.2**	28.5	<u>72.6</u>	<u>69.6</u>	<u>72.1</u>	<u>59.2</u>
Qwen2.5-VL-7B-Instruct (Bai et al., 2025)	-	46.2	25.0	67.5	63.1	64.6	53.3
+ Vanilla GRPO	2.1K (Geometry3K)	49.6	26.8	70.2	68.2	69.8	56.9
+ Ours	2.1K (Geometry3K)	<u>52.8</u>	<u>28.8</u>	72.3	69.2	**72.9**	<u>59.2</u>
+ Vanilla DAPO	2.1K (Geometry3K)	50.4	27.6	70.7	69.4	68.6	57.3
+ Ours	2.1K (Geometry3K)	51.9	**29.0**	72.2	**71.8**	71.4	**59.3**

The combination of both strategies achieves the best overall performance with an average accuracy of 59.3. This confirms that these two strategies provide complementary perspectives on samples, and their integration leads to more robust policy optimization.

4.3 Comparison with State-of-the-art approaches

In this subsection, we comprehensively compare our method with various state-of-the-art methods, including closed-source, open-source, and RLVR-related approaches. The experimental results are listed in Table 2. We can draw the following observations: (1) compared with those either distilled from large-scale chain-of-thought data or employing complex data augmentation strategies, our method, utilizing only 2.1k training samples, achieves comparable or even superior performance, significantly demonstrating our effectiveness. (2) Building upon both GRPO and DAPO, our strategy demonstrates promising performance gains. Specifically, we achieve more than 3.5% performance improvements, especially on the HallusionBench; our strategy achieves more than 5% improvements, further showing our effectiveness.

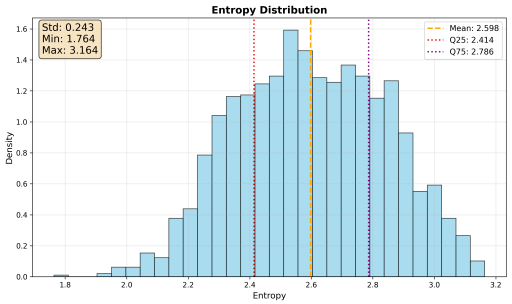


Figure 2: The distribution of pre-calculated entropy on the Geometry3K dataset, where x axis represents entropy, y axis denotes the probability density, Q25 and Q75 denotes 25th and 75th percentiles, respectively .

4.4 Hyper-parameter Sensitivity Analysis

In this section, we analyze the effects of hyperparameters, including both the number of perceptual and reasoning difficulty-based groups.

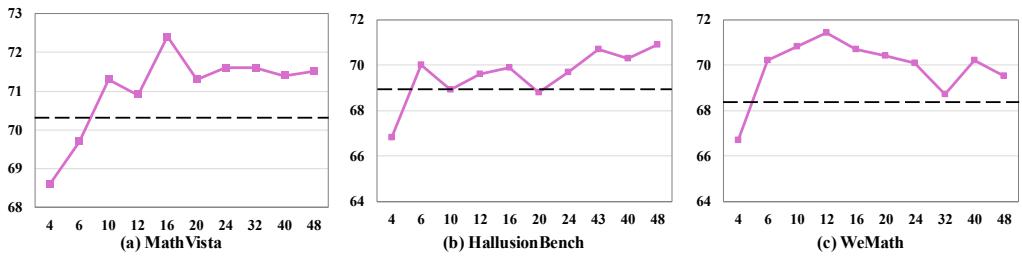


Figure 3: The Accuracy performance of different numbers of groups b on 3 representative benchmarks, where the x axis is the number of groups, and y axis is the results.

4.4.1 Groups under Perceptual difficulty-based strategy

As shown in Figure 2 , to regroup samples by entropy, we adopt the 25th and 75th percentiles as thresholds. This quantile-based choice is inherently distribution-aware, as it adapts to the empirical spread of entropy values rather than relying on arbitrary fixed cutoffs. Importantly, it produces a natural 1:2:1 partition of the data—approximately 25% easy, 50% medium, and 25% hard—which avoids the issue of overly sparse or overly dense categories. Such a balance is desirable for stable optimization: each group contains sufficient samples to provide reliable intra-group statistical estimates, while extremely low- and high-entropy cases are isolated rather than allowed to dominate normalization. Moreover, this three-level categorization is semantically interpretable, with low entropy corresponding to simple scenes, high entropy to complex ones, and the middle range capturing moderately difficult cases. Detailed cases representing the entropy of these three categories are illustrated in Appendix D.1.

4.4.2 Groups under Reasoning difficulty-based strategy

We report the effect of varying the number of groups for our reasoning-based strategy in Figure 3. We observe that the performance is relatively stable across a wide range of groups, suggesting that our method is robust to this hyperparameter. For instance, on MathVista and HallusionBench, the accuracy is steadily improved as the number of groups increases to around 12–16, after which the results plateau with only minor fluctuations. A similar trend is observed on WeMath, where the performance peaks at 12 groups but remains competitive even when more groups are introduced.

5 Related works

In this section, we overview of the related studies. Specifically, we first discuss representative strategy to construct multimodal reasoning models, including chain-of-thought distillation, reinforcement learning, and visual tool integration. Then we introduce the RLVR and its optimization variants, and finally, we highlight the key differences between ours and existing approaches.

5.1 Multimodal Reasoning Models

Chain-of-thought distillation. Supervised fine-tuning (SFT) on long chain-of-thought (CoT) data enables models to learn detailed reasoning traces, thereby improving reasoning accuracy. Specifically, building upon (Zhang et al., 2023), this strategy has proven effective through both transferring CoT-enhanced LLMs to multimodal settings (Du et al., 2025b) and training directly with multimodal reasoning data (Liu et al., 2023). Future works futher explore different forms of intermediate reasoning supervision (Dai et al., 2023; Yang et al., 2025c; Zhu et al., 2023).

Reinforcement learning. Another line of research leverages reinforcement learning (RL) to optimize reasoning trajectories beyond imitation. Most studies adopt PPO (Schulman et al., 2017) or GRPO (Shao et al., 2024b), with representative approaches such as (Zhang et al., 2025b; Shen et al., 2025; Wang et al., 2025b;c) that apply RL across diverse domains. We will elaborate on RLVR and GRPO in the following subsection(Section 5.2).

Visual tool integration. This paradigm moves beyond merely "thinking about images" toward actively querying, modifying, and generating visual information as intermediate steps in reasoning, forming a "visual chain of thought". The development of think-with-image can be roughly divided into three stages (Su et al., 2025) : from external tool exploration (Ma et al., 2024; Shao et al., 2024a; Ma et al., 2025), through programmatic manipulation (Surís et al., 2023; Fu et al., 2025), to intrinsic imagination (Zhao et al., 2025; Chen et al., 2025b). These three stages reflect interconnected capabilities—active exploration, structured reasoning, and generative planning—that together transform visual representations from static inputs into a dynamic workspace for thought.

5.2 Reinforcement Learning with Verifiable Reward (RLVR)

RLVR (Lambert et al., 2024) is an optimization paradigm that replaces subjective reward scores with verifiable signals. Its core algorithm, GRPO (Shao et al., 2024b), stabilizes training by comparing candidate responses within a group. Building on GRPO, subsequent studies can be broadly categorized into two directions: data-centric approaches, which expand the candidate and reward space through data manipulation or augmentation, and algorithm-centric approaches, which refine GRPO to strengthen semantic grounding and coherent reasoning.

Data-centric GRPO. This line of work enlarges the candidate set (Chen et al., 2025d) or restructures the training data (Chen et al., 2025c; Zhu et al., 2025) so that group comparisons capture richer behaviors. By manipulating data distributions (Zhu et al., 2025) or augmenting inputs (Li et al., 2025; Liu et al., 2025a), these methods expose models to a wider variety of responses, thereby increasing the likelihood of discovering high-quality verifiable signals.

Algorithm-centric GRPO. In contrast, algorithm-centric methods refine how verifiable signals guide reasoning. Rather than expanding candidate sets, they adapt GRPO to enhance semantic grounding (Yu et al., 2025a; Liu et al., 2025c) and logical coherence (Huang et al., 2025a; Wei et al., 2025). These approaches emphasize the role of visual grounding and promote reasoning chains where intermediate steps remain verifiable while supporting the final answer.

Difference. Compared with existing methods, we regroup the data in advantage calculation based on (1) the model's response uncertainty and (2) the entropy of image inputs when computing the *std*, and sharing the *std* within each group. This design prevents the model from overfitting to extreme samples and enhances its ability to capture the data distinction within each group.

6 Conclusion

In this work, we identify a critical challenge in GRPO-based reinforcement learning methods for multimodal reasoning tasks: **the *std*-based group normalization is sensitive to extreme samples**, such as response groups that are almost entirely positive or negative. While this issue exists in GRPO in general, it is significantly amplified in multimodal settings due to the joint influence of perceptual complexity and reasoning uncertainty.

To address this, we propose a simple yet effective **difficulty-aware re-grouping** strategy. By decomposing the sample difficulty into perceptual and reasoning aspects, we construct groups of samples with similar difficulty levels, allowing each group to share *std* during normalization. The normalized advantages with shared *std* from both aspects are combined via element-wise multiplication, effectively integrating data complexity and model uncertainty while preserving intra-group distinctions. By applying over GRPO and DAPO, our strategy achieves more than 2% average performance gains across multiple multimodal reasoning benchmarks.

Limitations and Future Work. While difficulty-aware regrouping significantly stabilizes GRPO-based multimodal reasoning, several directions remain open: (1) more precise difficulty estimation —— potentially learned from model behaviors to richer visual-semantic features —— could capture subtler perceptual and reasoning characteristics; (2) adaptive grouping strategies that dynamically adjust group sizes or composition as training progresses may better balance intra-group distinctions and mitigate extreme samples. Beyond technical refinements, the underlying principle of aligning optimization with sample difficulty also offers a general paradigm for stabilizing reinforcement learning optimization with multimodal inputs.

7 ETHICS STATEMENT

Our work focuses on algorithmic improvements for reinforcement learning-based multimodal reasoning tasks using publicly available datasets (e.g., MathVista, MathVerse, MathVision). We do not collect data from human subjects, and all datasets used are publicly released. We do not foresee any direct, immediate, or negative societal impacts stemming from the outcomes of our research.

8 REPRODUCIBILITY STATEMENT

To ensure reproducibility, we provide detailed descriptions of datasets, training configurations, and hyperparameters in both the main text and supplementary materials. We will provide our code to facilitate reproducibility.

REFERENCES

Anthropic. Claude 3.5 sonnet, 2024. URL https://www.anthropic.com/news/claude-3-5-sonnet.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *CoRR*, abs/2502.13923, 2025.

Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. SFT or rl? an early investigation into training r1-like reasoning large vision-language models. *CoRR*, abs/2504.11468, 2025a.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *CoRR*, abs/2505.09568, 2025b.

Mingrui Chen, Haogeng Liu, Hao Liang, Huaibo Huang, Wentao Zhang, and Ran He. Unlocking the potential of difficulty prior in rl-based multimodal reasoning. *CoRR*, abs/2505.13261, 2025c.

Xi Chen, Mingkang Zhu, Shaoteng Liu, Xiaoyang Wu, Xiaogang Xu, Yu Liu, Xiang Bai, and Hengshuang Zhao. Mico: Multi-image contrast for reinforcement visual reasoning. *CoRR*, abs/2506.22434, 2025d.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *CoRR*, abs/2412.05271, 2024.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*, 2023.

DeepSeek-AI, 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. *CoRR*, abs/2501.12948, 2025.

Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. Openvlthinker: An early exploration to complex vision-language reasoning via iterative self-improvement. *CoRR*, abs/2503.17352, 2025.

Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *CoRR*, abs/2504.07491, 2025a.

Yifan Du, Zikang Liu, Yifan Li, Wayne Xin Zhao, Yuqi Huo, Bingning Wang, Weipeng Chen, Zheng Liu, Zhongyuan Wang, and Ji-Rong Wen. Virgo: A preliminary exploration on reproducing o1-like MLLM. *CoRR*, abs/2501.01904, 2025b.

Xingyu Fu, Minqian Liu, Zhengyuan Yang, John Corring, Yijuan Lu, Jianwei Yang, Dan Roth, Dinei Florencio, and Cha Zhang. Refocus: Visual editing as a chain of thought for structured image understanding. *CoRR*, abs/2501.05452, 2025.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *CVPR*, 2024.

Lujun Gui and Qingnan Ren. Training reasoning model with dynamic advantage estimation on reinforcement learning. https://www.notion.so/Training-Reasoning-Model-with-Dynamic-Advantage-Estimation-on-Reinforcement-Learning-1a830cc0904681fa9df3e076b6557a3e, 2025. Notion Blog.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *CoRR*, abs/2503.06749, 2025a.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *CoRR*, abs/2503.06749, 2025b.

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

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. Tülu 3: Pushing frontiers in open language model post-training. *CoRR*, abs/2411.15124, 2024.

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. *CoRR*, abs/2408.03326, 2024.

Yuting Li, Lai Wei, Kaipeng Zheng, Jingyuan Huang, Linghe Kong, Lichao Sun, and Weiran Huang. Vision matters: Simple visual perturbations can boost multimodal math reasoning. *CoRR*, abs/2506.09736, 2025.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 36: 34892–34916, 2023.

Xiangyan Liu, Jinjie Ni, Zijian Wu, Chao Du, Longxu Dou, Haonan Wang, Tianyu Pang, and Michael Qizhe Shieh. Noisyrollout: Reinforcing visual reasoning with data augmentation. *CoRR*, abs/2504.13055, 2025a.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. *CoRR*, abs/2503.20783, 2025b.

Ziyu Liu, Yuhang Zang, Yushan Zou, Zijian Liang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual agentic reinforcement fine-tuning. *CoRR*, abs/2505.14246, 2025c.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In *ACL*, 2021.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *ICLR*, 2024.

Ruilin Luo, Zhuofan Zheng, Yifan Wang, Yiyao Yu, Xinzhe Ni, Zicheng Lin, Jin Zeng, and Yujiu Yang. URSA: understanding and verifying chain-of-thought reasoning in multimodal mathematics. *CoRR*, abs/2501.04686, 2025.

Yan Ma, Linge Du, Xuyang Shen, Shaoxiang Chen, Pengfei Li, Qibing Ren, Lizhuang Ma, Yuchao Dai, Pengfei Liu, and Junjie Yan. One rl to see them all: Visual triple unified reinforcement learning. *CoRR*, abs/2505.18129, 2025.

Zixian Ma, Jianguo Zhang, Zhiwei Liu, Jieyu Zhang, Juntao Tan, Manli Shu, Juan Carlos Niebles, Shelby Heinecke, Huan Wang, Caiming Xiong, Ranjay Krishna, and Silvio Savarese. TACO: learning multi-modal action models with synthetic chains-of-thought-and-action. *CoRR*, abs/2412.05479, 2024.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, Ping Luo, Yu Qiao, Qiaosheng Zhang, and Wenqi Shao. Mm-eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning. *CoRR*, abs/2503.07365, 2025.

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Jiapeng Wang, Zhuoma Gongque, Shanglin Lei, Yifan Zhang, Zhe Wei, Miaoxuan Zhang, Runfeng Qiao, Xiao Zong, Yida Xu, Peiqing Yang, Zhimin Bao, Muxi Diao, Chen Li, and Honggang Zhang. We-math: Does your large multimodal model achieve human-like mathematical reasoning? In *ACL*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017.

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

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In *NeurIPS*, 2024a.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300, 2024b.

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, Ruochen Xu, and Tiancheng Zhao. VLM-R1: A stable and generalizable r1-style large vision-language model. *CoRR*, abs/2504.07615, 2025.

Zhaochen Su, Peng Xia, Hangyu Guo, Zhenhua Liu, Yan Ma, Xiaoye Qu, Jiaqi Liu, Yanshu Li, Kaide Zeng, Zhengyuan Yang, et al. Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers. *CoRR*, abs/2506.23918, 2025.

Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. In *ICCV*, 2023.

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *CoRR*, abs/2504.08837, 2025a.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. In *NeurIPS*, 2024.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *CoRR*, abs/2508.18265, 2025b.

Xiyao Wang, Chunyuan Li, Jianwei Yang, Kai Zhang, Bo Liu, Tianyi Xiong, and Furong Huang. Llava-critic-r1: Your critic model is secretly a strong policy model. *CoRR*, abs/2509.00676, 2025c.

Xiyao Wang, Zhengyuan Yang, Chao Feng, Hongjin Lu, Linjie Li, Chung-Ching Lin, Kevin Lin, Furong Huang, and Lijuan Wang. Sota with less: Mcts-guided sample selection for data-efficient visual reasoning self-improvement. *CoRR*, abs/2504.07934, 2025d.

Lai Wei, Yuting Li, Kaipeng Zheng, Chen Wang, Yue Wang, Linghe Kong, Lichao Sun, and Weiran Huang. Advancing multimodal reasoning via reinforcement learning with cold start. *CoRR*, abs/2505.22334, 2025.

Zijian Wu, Jinjie Ni, Xiangyan Liu, Zichen Liu, Hang Yan, and Michael Qizhe Shieh. Synthrl: Scaling visual reasoning with verifiable data synthesis. *CoRR*, abs/2506.02096, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *CoRR*, abs/2505.09388, 2025a.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, Bo Zhang, and Wei Chen. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *CoRR*, abs/2503.10615, 2025b.

Yue Yang, Ajay Patel, Matt Deitke, Tanmay Gupta, Luca Weihs, Andrew Head, Mark Yatskar, Chris Callison-Burch, Ranjay Krishna, Aniruddha Kembhavi, and Christopher Clark. Scaling text-rich image understanding via code-guided synthetic multimodal data generation. *CoRR*, abs/2502.14846, 2025c.

Huanjin Yao, Jiaxing Huang, Wenhao Wu, Jingyi Zhang, Yibo Wang, Shunyu Liu, Yingjie Wang, Yuxin Song, Haocheng Feng, Li Shen, and Dacheng Tao. Mulberry: Empowering MLLM with o1-like reasoning and reflection via collective monte carlo tree search. *CoRR*, abs/2412.18319, 2024.

En Yu, Kangheng Lin, Liang Zhao, Jisheng Yin, Yana Wei, Yuang Peng, Haoran Wei, Jianjian Sun, Chunrui Han, Zheng Ge, Xiangyu Zhang, Daxin Jiang, Jingyu Wang, and Wenbing Tao. Perception-r1: Pioneering perception policy with reinforcement learning. *CoRR*, abs/2504.07954, 2025a.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Weinan Dai, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. DAPO: an open-source LLM reinforcement learning system at scale. *CoRR*, abs/2503.14476, 2025b.

Tianyu Yu, Bo Ji, Shouli Wang, Shu Yao, Zefan Wang, Ganqu Cui, Lifan Yuan, Ning Ding, Yuan Yao, Zhiyuan Liu, Maosong Sun, and Tat-Seng Chua. RLPR: extrapolating RLVR to general domains without verifiers. *CoRR*, abs/2506.18254, 2025c.

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. *CoRR*, abs/2503.12937, 2025a.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, Peng Gao, and Hongsheng Li. MATHVERSE: does your multi-modal LLM truly see the diagrams in visual math problems? In *ECCV*, 2024.

Yifan Zhang, Xingyu Lu, Xiao Hu, Chaoyou Fu, Bin Wen, Tianke Zhang, Changyi Liu, Kaiyu Jiang, Kaibing Chen, Kaiyu Tang, Haojie Ding, Jiankang Chen, Fan Yang, Zhang Zhang, Tingting Gao, and Liang Wang. R1-reward: Training multimodal reward model through stable reinforcement learning. *CoRR*, abs/2505.02835, 2025b.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *CoRR*, abs/2302.00923, 2023.

Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, Ankur Handa, Ming-Yu Liu, Donglai Xiang, Gordon Wetzstein, and Tsung-Yi Lin. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. *CoRR*, abs/2503.22020, 2025.

Yaowei Zheng, Junting Lu, Shenzhi Wang, Zhangchi Feng, Dongdong Kuang, and Yuwen Xiong. Easyr1: An efficient, scalable, multi-modality rl training framework. https://github.com/hiyouga/EasyR1, 2025.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *CoRR*, abs/2304.10592, 2023.

Linghao Zhu, Yiran Guan, Dingkang Liang, Jianzhong Ju, Zhenbo Luo, Bin Qin, Jian Luan, Yuliang Liu, and Xiang Bai. Shuffle-r1: Efficient RL framework for multimodal large language models via data-centric dynamic shuffle. *CoRR*, abs/2508.05612, 2025.

A The Use of Large Language Models(LLMs)

We conducted a study on improving GRPO to further enhance the reasoning capability of MLLMs, achieving substantial performance gains on datasets such as MathVerse (Zhang et al., 2024), MathVision (Wang et al., 2024), and MathVista (Lu et al., 2024). During the preparation of this manuscript, we used LLMs to assist with tasks such as grammar correction, language refinement, and logical checking. However, we confirm that no outputs from the LLMs were directly used; instead, all content underwent careful verification and reconstruction by the authors.

B Prompt Design

We use a "Thinking prompt" to formalize the output of the model. It requires the model to put its reasoning process within <think>...</think> and the final answer in `\boxed{}`. We keep the system prompt of Qwen2.5-VL (Bai et al., 2025) and prepend the "Thinking prompt" to the user message. The same format is used for both training and evaluation. The full instruction prompt is as follows:

Prompt Example

SYSTEM:
You are a helpful assistant.
USER:
You FIRST think about the reasoning process as an internal monologue and then provide the final answer. The reasoning process MUST BE enclosed within <think> </think> tags. The final answer MUST BE put in `\boxed{}`.<QUESTION>

C Experiment Settings

Reward Calculation. We adopt a combination of format reward and accuracy reward as the final reinforcement learning signal. The two components are defined as follows:

$$r_{\text{format}} = \begin{cases} 1, & \text{if the output format is correct,} \\ 0, & \text{otherwise,} \end{cases} \tag{21}$$

$$r_{\text{acc}} = \begin{cases} 1, & \text{if the answer matches the ground truth,} \\ 0, & \text{otherwise.} \end{cases} \tag{22}$$

The overall reward is computed as the weighted sum of the two:

$$r_{\text{overall}} = 0.1 \times r_{\text{format}} + 0.9 \times r_{\text{acc}}. \tag{23}$$

A smaller weight is assigned to the format reward, since response formatting is relatively easy to learn compared with accuracy.

D CASE STUDY

D.1 PERCEPTUAL DIFFICULTY-BASED RE-GROUPING CASES

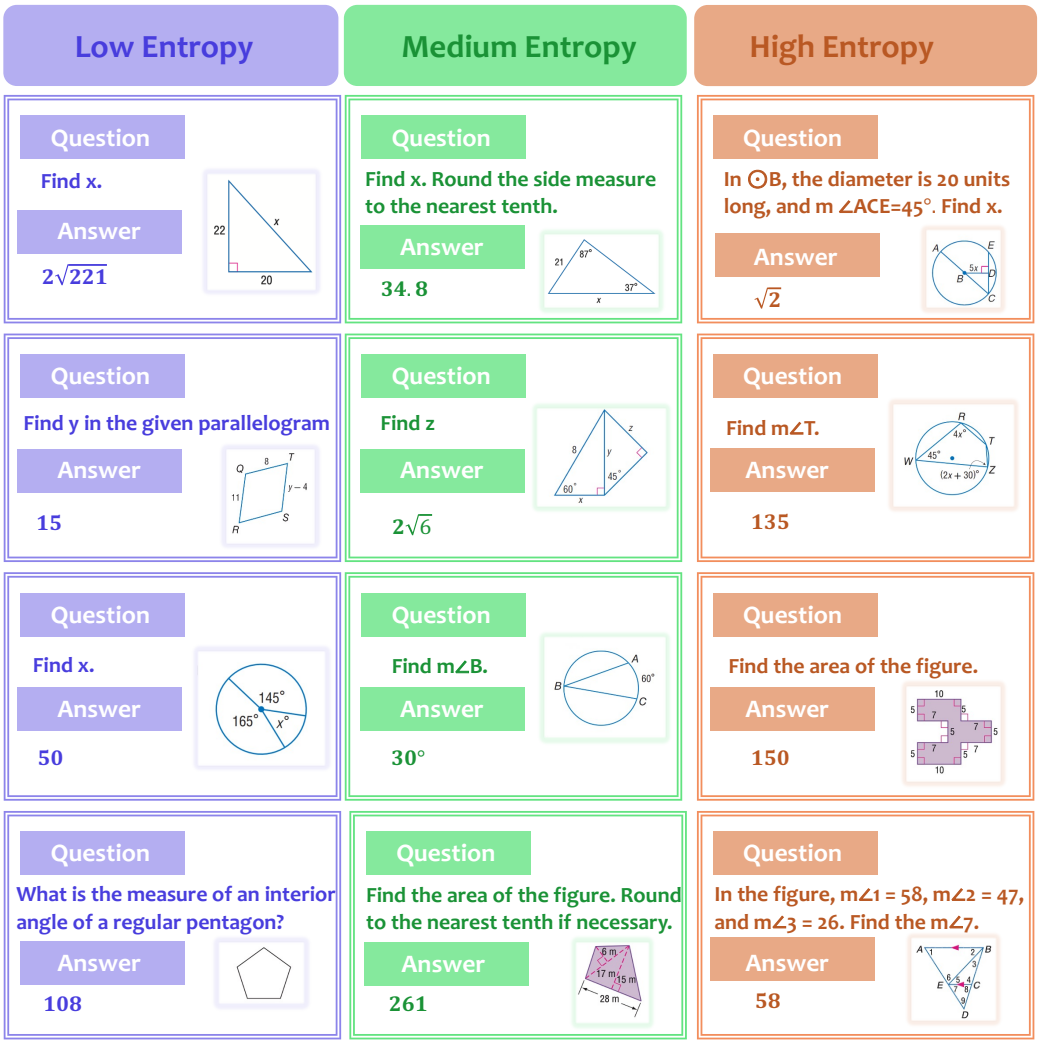
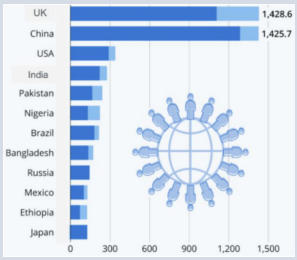


Figure 4: Illustrative examples of different levels of entropy

D.2 Demonstration of improved perception and reasoning capabilities



Question:
According to the chart, does China have the second largest population in the world? You final answer can only be \boxed{yes} or \boxed{no}.

Source: HallusionBench
Problem version: chart
figure id: 3_2
question_id: 3

Ours:

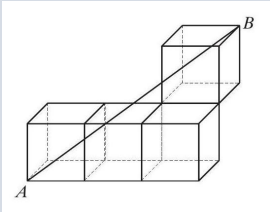
<think>
The chart represents population data for various countries, with China and the UK having the highest populations at 1,425.7 and 1,428.6 million respectively.
Since the chart indicates China's population is slightly lower than the UK's, it suggests that China does indeed have the second largest population in the world according to this data.
</think>
\boxed{yes}

Vanilla GRPO:

<think>
The chart shows the population of various countries. The population of China is listed as 1,425.7 million, and the population of the UK is listed as 1,428.6 million. Since the UK has a slightly higher population, China does not have the second largest population in the world. The second largest population in the world is the UK.
</think>
\boxed{no}

Figure 5: Case Study1 showing improved reasoning capability over vanilla GRPO

Question:
Each of the cubes in the figure has the length of an edge equal to 1. What is the length of the segment AB?
(A) $\sqrt{17}$ (B) 7 (C) $\sqrt{13}$ (D) $\sqrt{7}$ (E) $\sqrt{14}$



Source: Mathvision
question_id: 209

Ours:

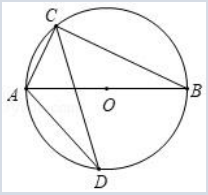
<think>
To find the length of the segment AB, we can use the three-dimensional distance formula, which is an extension of the Pythagorean theorem. The distance d between two points (x_1, y_1, z_1) and (x_2, y_2, z_2) in three-dimensional space is given by: $d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}$
Here, point A is at (0, 0, 0) and point B is at (3, 2, 2). Plugging these coordinates into the formula, we get:
$AB = \sqrt{(3-0)^2 + (2-0)^2 + (2-0)^2} = \sqrt{3^2 + 2^2 + 2^2} = \sqrt{17}$
Therefore, the length of the segment AB is $\sqrt{17}$.
</think>
\boxed{A}

Vanilla GRPO:

<think>
To find the length of the segment AB, we need to consider the three-dimensional space formed by the cubes. The segment AB can be visualized as a diagonal that spans across three unit cubes in the x, y, and z directions.
1. The segment AB moves 3 units in the x-direction, 2 units in the y-direction, and 1 unit in the z-direction.
2. The length of AB can be found using the three-dimensional distance formula, which is similar to the Pythagorean theorem but extended to three dimensions: $d = \sqrt{x^2 + y^2 + z^2}$
3. Substituting the values x = 3, y = 2, and z = 1: $AB = \sqrt{3^2 + 2^2 + 1^2} = \sqrt{14}$
Therefore, the length of the segment AB is $\sqrt{14}$
</think>
\boxed{$\sqrt{14}$}

Figure 6: Case Study2 showing improved reasoning capability over vanilla GRPO

Question:
As shown in the figure, AB is the diameter of ⊙ O, CD is the chord of ⊙ O, ∠ADC = 26.0, then the degree of ∠CAB is ():
(A) 26° (B) 74° (C) 64° (D) 54°



Source: Mathvista
question_id: 59

Ours:

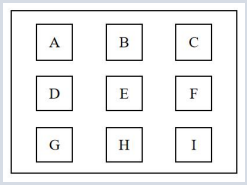
<think>
Since AB is the diameter of circle O, angle ACB is a right angle (90°) because any angle inscribed in a semicircle is a right angle. We are given that angle ADC is 26°. Because angles subtended by the same arc are equal, angle ACB (which is 90°) and angle ADB (which is also 26°) are related to the arc AB. However, to find angle CAB, we can use the fact that the sum of angles in triangle ACB is 180°. Since angle ACB is 90° and angle ABC (opposite to side AC) is the same as angle ADC (26°) due to the inscribed angle theorem, we can calculate angle CAB as follows:
angle CAB = 180° - angle ACB - angle ABC
angle CAB = 180° - 90° - 26°
angle CAB = 64°
Therefore, the degree of angle CAB is 64°.
</think>
\boxed{C}

Vanilla GRPO:

<think>
Since AB is the diameter of circle O, angle ACB is a right angle (90 degrees) because the angle subtended by a diameter in a semicircle is always 90 degrees. Given that angle ADC is 26 degrees, and angles ADC and CAB are subtended by the same arc AC, they are equal. Therefore, angle CAB is also 26 degrees.
So, the degree of angle CAB is 26 degrees.
</think>
\boxed{A}

Figure 7: Case Study3 showing improved reasoning capability over vanilla GRPO

Question:
As shown in the diagram, in a building, it is known that A lives to the left of B. Who lives below D?
A. A; B. E; C. F; D. G; E. No correct answer



Source: Wemath
question_id: 2steps_17@2steps_1

Ours:

<think>
The diagram shows a grid with letters A through I. The question states that A lives to the left of B. This means that if A is in a certain position, B must be to the right of A in the same row. However, the question asks who lives below D. Since D is in the second row, the person below D would be in the third row, which is G.
So, the person who lives below D is G.
</think>
\boxed{D}

Vanilla GRPO:

<think>
The diagram shows a grid with letters A through I. The question states that A lives to the left of B. This means that if we look at the horizontal row where A is located, B must be to the right of A. Now, we need to determine who lives below D.
In the grid:
- D is in the second row, first column.
- The letter below D is E.
Therefore, the person who lives below D is E.
</think>
\boxed{B}

Figure 8: Case Study4 showing improved reasoning capability over vanilla GRPO