MAPO: Mixed Advantage Policy Optimization

Anonymous authors
Paper under double-blind review

Abstract

Recent advances in reinforcement learning for foundation models, such as Group Relative Policy Optimization (GRPO), have significantly improved the performance of foundation models on reasoning tasks. Notably, the advantage function serves as a central mechanism in GRPO for ranking the trajectory importance. However, existing explorations encounter both advantage reversion and advantage mirror problems, which hinder the reasonable advantage allocation across different query samples. In this work, we propose an easy but effective GRPO strategy, **M**ixed **A**dvantage **P**olicy **O**ptimization (**MAPO**). We reveal that the trajectory appears with different certainty and propose the advantage percent deviation for samples with high-certainty trajectories. Furthermore, we dynamically reweight the advantage function for samples with varying trajectory certainty, thereby adaptively configuring the advantage function to account for sample-specific characteristics. Comparison with related state-of-the-art methods, along with ablation studies on different advantage variants, validates the effectiveness of our approach.

1 Introduction

Recent advances in the reasoning capabilities of Foundation Model (FM) Jaech et al. (2024); Team et al. (2025); Guo et al. (2025); Wen et al. (2025); Guo et al. (2025); Wang et al. (2025b) have been largely driven by improvements in long Chain of Thought (CoT) generation. Among various enhancement strategies, Reinforcement Learning (RL) Ouyang et al. (2022); OpenAI (2023); Shao et al. (2024); Hu et al. (2025); Xiong et al. (2025) has emerged as a powerful post-training technique, enabling FM to refine their CoT reasoning through self-improvement. Therefore, RL serves as the key mechanism for unlocking the reasoning ability in various domains.

Notably, Group Relative Policy Optimization (GRPO) Shao et al. (2024) is introduced as a popular reinforcement strategy. GRPO generates and refines a group of reasoning paths through the group-relative advantage estimation based on rule-based reward functions. Thus, a key difference with traditional reinforcement methods, such as proximal policy optimization Schulman et al. (2017) and direct preference optimization Rafailov et al. (2023); Chen et al. (2024); Liu et al. (2025a), is that GRPO eliminates the need for an additional learned reward critic model, instead leveraging efficient sampling from the Foundation Model policy. Witnessing the success of Group Relative Policy Optimization, its advantage function plays a key role in promoting trajectories with relatively higher advantages, thereby guiding the policy model to update towards more reliable directions. Despite recent advancements, GRPO and its variants generally maintain a fixed advantage formulation throughout the entire training cycle Guo et al. (2025); Yao et al. (2025); Guo et al. (2025). However, this approach overlooks a significant challenge: *the fixed advantage fails to provide meaningful signals for samples with varying trajectory certainty degrees*.

To analyze the drawbacks of existing advantage formulations, we first define ***trajectory certainty*** within the sampling group. The advantage is computed from verifiable rewards, typically format and accuracy metrics, to jointly measure the trajectory score Guo et al. (2025); Shao et al. (2024); Yao et al. (2025); Zhang et al. (2025); Liu et al. (2025b); Xu & Ding (2025). For a sampled trajectory, we declare the **success** only if it achieves the correct answer on all reward metrics. Consequently, each trajectory can be viewed as a Bernoulli trial with outcome: failure or success. Then, in the group sampling, the number of successes over repeated draws follows a binomial distribution, and *high-certainty samples tend to yield nearly identical outcomes across draws*, *i.e.*, samples that are too hard or too easy. We then formally derive the definition of trajectory certainty as follows:

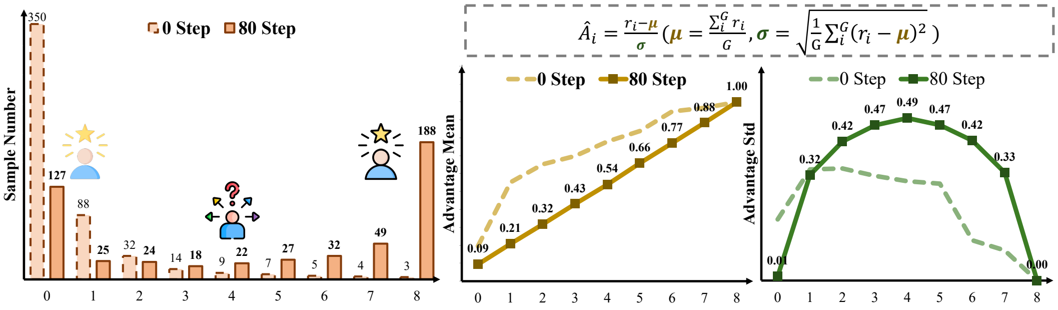


Figure 1: **Observation**. During the reinforcement, different samples appear with diverse successful trajectory numbers $N = \sum_{i=1}^{G} \mathbf{1}_{\{r_i=1\}}$ (**X-axis**). Samples with lowest trajectory certainty (?) tends to achieve most diverse prediction pattern, *i.e.*, $N = 4$. Experiments are conducted on the Geo3K with rollout number $G = 8$.

> ***Trajectory Certainty in GRPO**: High certainty ☆ corresponds to trajectories with lower prediction variance, while low certainty ? reflects higher variance.*

We analyze the sample behavior in fig. 1 and reveal two underlying limitations of the existing advantage paradigm. *First*, ***Advantage Reversion***: high-certainty samples may receive more differentiated advantage allocations than low-certainty ones. Specifically, a high-certainty sample (☆) with rewards $\boldsymbol{r}_{\text{High}} = \{0.9, 1.0, 1.0, 1.0\}$ receives a more discriminative advantage allocation than a low-certainty sample (?) with $\boldsymbol{r}_{\text{Low}} = \{0.1, 0.9, 1.0, 1.0\}$ due to **a small advantage standard deviation** σ. However, high-certainty samples do not require strong penalization, whereas low-certainty trajectories benefit from stronger correction. *Second*, ***Advantage Mirror***: high-certainty samples (☆ & ☆) also require distinct advantage allocation for extreme cases. In particular, the existing advantage formulation does not take into account the **monotonic advantage scores** μ and therefore treats easy and hard samples indistinguishably. The core issue is that the same advantage formulation cannot be applied uniformly across samples with different trajectory certainty. In summary, this motivates us to rethink the advantage design and decomposes it into two sub-questions: **i)** *how to design the advantage function for high-certainty samples*? and **ii)** *how to adaptively combine advantage functions for samples with varying trajectory certainty*?

To address the question **i)**, we introduce the Advantage Percent Deviation (APD), which replaces the advantage from standard z-score normalization to relative normalization. Specifically, the original advantage formulation is expressed as $\hat{A}_i = \frac{r_i - \mu}{\sigma}$. For high-certainty sample trajectories, this formulation fails to capture the overall level of reward scores. Besides, variance in the rollout trajectory can yield a small $\sigma = \text{std}(\boldsymbol{r})$, which in turn leads to numerical instability and uncontrollable boundary on advantage allocation. To deal with this drawback, we introduce a novel advantage function for high-certainty samples as $\hat{A}_i^{APD} = \frac{r_i - \mu}{\mu}$. Regarding question **ii)**, we propose the Trajectory Certainty Reweight (TCR) to determine the sample advantage function based on trajectory certainty. Inspired by Bernoulli sampling, each trajectory is treated as either a success or a failure. A trajectory group exhibits the highest uncertainty when the success-to-failure ratio is fifty percent. Therefore, we use trajectory certainty to dynamically reweight the advantage function from $\hat{A}_i$ to $\hat{A}_i^{APD}$. In this work, we argue that the existing advantage formulation is not consistently appropriate for samples with varying levels of trajectory certainty. To address this issue, we propose a simple yet effective method, **M**ixed **A**dvantage **P**olicy **O**ptimization (**MAPO**), which rethinks the advantage formulation in GRPO. To validate our approach, we conduct extensive experiments across multiple datasets using the Qwen2.5-VL-7B architecture, demonstrating the superior performance in both In-Domain and Out-of-Domain aspects. Our contributions are summarized as follows:

- We focus on the Group Relative Policy Optimization paradigm and reveal that existing advantage formulation faces two unavoidable challenges: advantage reversion and advantage mirror.
- We propose Mixed Advantage Policy Optimization (MAPO), a simple yet effective method to overcome existing advantage limitations. Preliminary, we introduce trajectory certainty to evaluate sample behavior. We propose Advantage Percent Deviation for high-certainty advantage estimation and utilize Trajectory Certainty Reweight to dynamically construct the advantage function.

- We perform a comprehensive analysis on reasoning scenarios, including mathematics and emotion fields. Through a series of ablation studies, the promising results empirically validate the effectiveness of the proposed mixture advantage strategies in enhancing GRPO overall performance.

2 Related Works

2.1 Foundation Model

The development of Large Language Model (LLM) has revolutionized artificial intelligence, significantly transforming the way machines understand and generate human language. Notable examples of LLM include the GPT series Radford et al. (2019); Brown et al. (2020); OpenAI (2023), Meta LLaMA Touvron et al. (2023), and Google PaLM Chowdhery et al. (2022); Anil et al. (2023), all of which have demonstrated impressive capabilities in natural language understanding and generation. These advancements have sparked considerable interest in extending LLM to handle multi-modal inputs, particularly by incorporating vision components, which has led to the development of Multimodal Large Language Model (MLLM). Building on the success of LLM, growing interest has emerged in constructing end-to-end Multimodal Large Language Model (MLLM) systems, such as Flamingo Alayrac et al. (2022), BLIP-2 Li et al. (2022; 2023), InstructBLIP Dai et al. (2023), QWen-VL Bai et al. (2023b), LLaVA Liu et al. (2023b;a); Zhu et al. (2024); Li et al. (2024), and VILA Lin et al. (2023); Fang et al. (2024); Liu et al. (2025d). Existing MLLM solutions typically rely on visual extractors Radford et al. (2021); Dosovitskiy et al. (2021); Caron et al. (2021) to encode visual features, using a connector module to project visual tokens into the word embedding space of the LLM, *i.e.*, treating visual input as a foreign language Wang et al. (2023). Subsequently, the visual and textual tokens are concatenated and fed into the LLM. The LLM is then used to perform various vision-language tasks in an auto-regressive manner. As a result, foundation models are gradually evolving from a single-textual modality to multimodal capabilities. However, existing works predominantly focus on supervised fine-tuning (SFT) on large-scale pre-training datasets. The success of OpenAI o1 Radford et al. (2019); Jaech et al. (2024); Ouyang et al. (2022) highlights the powerful potential of reinforcement learning in post-training to enhance model reasoning capabilities. With the open-sourcing of Deepseek-R1 Guo et al. (2025) and Qwen Bai et al. (2023a;b); Yang et al. (2025b;a), reasoning models are now widely deployed locally, drawing attention from the research community to the efficiency of long chain-of-thought generation for foundation models. And utilizing the reinforcement technique to empower foundation models with reasoning capabilities has emerged as a pivotal methodology beyond the limitations of SFT.

2.2 Group Relative Policy Optimization

Several methods have been proposed to elicit reasoning abilities on mathematical and scientific problems, enabling foundation models to better handle inference and analysis. Especially, Group Relative Policy Optimization (GRPO) has recently garnered significant attention in the research field, as its rule-based reward function effectively enhances the reasoning capabilities of large models. Existing exploration or variants of GRPO could normally be divided into the following streams. ❶ *Think Trajectory Diversity*. This paradigm focuses on diversifying the thinking process to facilitate a more meaningful candidate rollout. Specifically, it boosts the trajectory from two angles: input perturbation and process polish. First, constructing the data augmentation technique for Multimodal Large Language Model to enhance both the quantity and quality of training data. NoisyRollout Liu et al. (2025b) leverages the noise annealing schedule to construct the noisy image text pairs. VP Li et al. (2025) introduces three targeted perturbations: distractor concatenation, dominance-preserving mixup, and random rotation. Share-GRPO Yao et al. (2025) turns to expand the question space for a given question via data transformation. Second, polishing the thinking process acts as a reliable direction to monitor the thinking behavior. StepGRPO Zhang et al. (2025) requires the think process to explicitly reveal key intermediate steps. Both SophiaVL-R1 Fan et al. (2025) and GRPO-CARE Chen et al. (2025b) utilize an external thinking reward model that evaluates the quality of the entire thinking process. MGRPO Ding et al. (2025) recycles previous think messages for self-correction learning. Hint-GRPO Huang et al. (2025) adaptively provides hints to the samples. However, the aforementioned solutions require constructing dedicated data augmentation strategies or modifying the thinking process, which introduces additional computational costs or an external thinking reward evaluation model. ❷ *Reward Formulation Refinement*. With respect to verifiable reward

Table 1: **Weakness** for different GRPO variants. Refer to Sec. 2.2 for details.

Methods	Input Space Augmentation	Think Cost Increase	Specific Task Adaption	Additional Hyper-Parameter
Think Trajectory Diversity				
NoisyRollout	✓ (Noisy Distortion)			✓ (Initial Noise Strength)
VP	✓ (Visual Augmentation)			✓ (Perturbation Types)
Share-GRPO	✓ (Textual Enrichment)			✓ (Question Variants Number)
StepGRPO		✓ (Step Think)		✓ (Key Steps Number)
GRPO-CARE		✓ (Reference Model)		✓ (Consistency Coefficient)
Reward Formulation Refinement				
Visual-RFT			✓ (Visual IoU Reward)	
GRPO-λ			✓ (Length Penalty)	✓ (Top-λ Fraction)
GRPO-LEAD			✓ (Length Reward)	✓ (Advantage Rescale Factor)
Advantage Estimation Redesign				
SEED-GRPO				✓ (Advantage Rescale Factor)
GPG				✓ (Valid Sample Threshold)

construction Bi et al. (2024); Team et al. (2025), it is the predefined rules and normally incorporates the Format Reward and Accuracy Reward. The former requires the model output should meet the required HTML tag format of `<think>` and `<answer>`. The latter is determined by comparing the model output class with the ground truth class, yielding a value of 1 for correct classification and 0 for incorrect classification. Thus, recent works design different verifiable reward functions for different specific tasks. For instance, Visual-RFT Liu et al. (2025f) proposes the intersection over union reward for object detection. VisionReasoner Liu et al. (2025c) introduces diverse perception rewards in a unified framework. Both GRPO-λ Dai et al. (2025a) and GRPO-LEAD Zhang & Zuo (2025) consider the over-length penalty reward. However, this pattern typically focuses on adapting to specific tasks, which limits cross-task generalization. Additionally, it requires careful tuning of hyperparameter reward weights for different reward metrics. Therefore, this pattern fails to achieve robust performance across diverse real-world application settings. ❸ *Advantage Estimation Redesign*. Towards advantage estimation, recent researches investigate better trajectory importance measurement via reformulation or rescaling operation. Dr. GRPO Liu et al. (2025e) and GPG Chu et al. (2025b) consider removing the standard deviation to alleviate the reward bias. SEED-GRPO Chen et al. (2025a) reweights the advantages based on the semantic entropy to measure the output uncertainty. KRPO Wang et al. (2025a) introduces a lightweight Kalman filter approach for accurate advantage estimation. But this paradigm faces the hyperparameter selection dilemma, or consistent advantages for different samples. We conclude the weakness of existing GRPO variants in Tab. 1. In our work, we reveal that *samples appear distinct trajectory certainty behavior and utilizing uniform advantage strategy unavoidably degrades partial samples optimization*. Therefore, we dynamically set the advantage function based on the trajectory certainty to boost the overall reinforcement effect.

3 Methodology

3.1 Preliminary

Group Relative Policy Optimization (GRPO) Shao et al. (2024) is a variant of Proximal Policy Optimization (PPO) Schulman et al. (2017) originally developed to enhance mathematical reasoning in LLM. However, it can also be effectively adapted to improve visual reasoning in Multimodal Large Language Model. GRPO begins by constructing the current policy model π_θ and a reference model π_{old}, where the latter represents the “old” policy or the policy from a previous iteration. Let ρ_Q denote the distribution of prompts or questions. Given a prompt $q \sim \rho_Q$, GRPO samples a group of outputs $o_1, o_2, \ldots, o_G$ from the old model π_{old}. It then optimizes the policy model π_θ by maximizing the following objective function:

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

where β is the hyper-parameter. $f_\epsilon(x, y) = \min(xy, \text{clip}(x, 1-\epsilon, 1+\epsilon)y)$. $\hat{A}_i$ is the advantage calculated based on the relative rewards of the outputs inside each group. To be precise, for each question q, a group of outputs $\{o_1, o_2, \ldots, o_G\}$ are sampled from the old policy model π_{old}. A reward function (R) is then used to score the outputs, yielding G rewards $\boldsymbol{r} = \{r_1, r_2, \ldots, r_G\}$

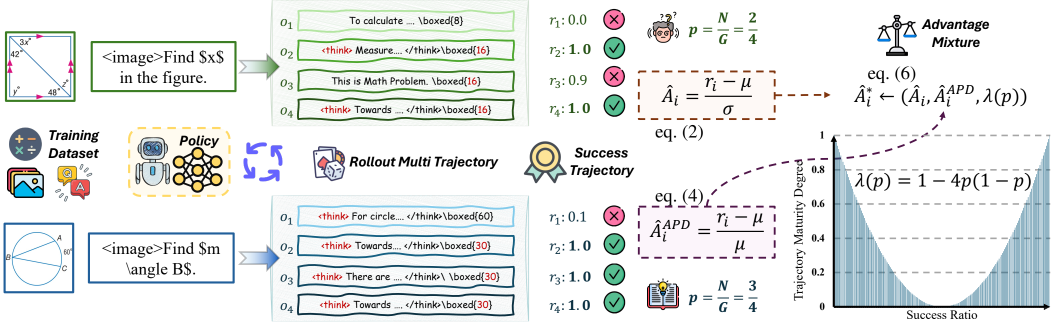

3x°
42°
y°
48°
z°
<image>Find x in the figure.
o_1 To calculate .... \boxed{8}
o_2 <think> Measure.... </think>\boxed{16}
o_3 This is Math Problem. \boxed{16}
o_4 <think> Towards .... </think>\boxed{16}
$r_1: 0.0$
$r_2: \mathbf{1.0}$
$r_3: 0.9$
$r_4: \mathbf{1.0}$
$p = \frac{N}{G} = \frac{2}{4}$
$\hat{A}_i = \frac{r_i - \mu}{\sigma}$
eq. (2)
Advantage Mixture
eq. (6)
$\hat{A}_i^* \leftarrow (\hat{A}_i, \hat{A}_i^{APD}, \lambda(p))$
Training Dataset
Policy
Rollout Multi Trajectory
Success Trajectory
eq. (4)
$\hat{A}_i^{APD} = \frac{r_i - \mu}{\mu}$
A
B
C
60°
<image>Find $m \angle B$.
o_1 <think> For circle.... </think>\boxed{60}
o_2 <think> Towards.... </think>\boxed{30}
o_3 <think> There are .... </think>\ \boxed{30}
o_4 <think> Towards .... </think>\boxed{30}
$r_1: 0.1$
$r_2: \mathbf{1.0}$
$r_3: \mathbf{1.0}$
$r_4: \mathbf{1.0}$
$p = \frac{N}{G} = \frac{3}{4}$
$\lambda(p) = 1 - 4p(1-p)$
Trajectory Maturity Degree
Success Ratio
0
0.2
0.4
0.6
0.8
1

Figure 2: **Architecture illustration of MAPO**. We reveal that the trajectory certainty varies across samples. In general, we introduce the Advantage Percent Deviation to replace the advantage function for high-certainty elements. We utilize the Trajectory Certainty Reweight to dynamically reweight the advantage function via trajectory certainty. Assume rollout number $G = 4$. Best viewed in color. Zoom in for details. See Sec. 3.2.

correspondingly, where $r_i = R(q, o_i)$. The mean reward is then calculated as $\mu = \frac{1}{G}\sum_{i=1}^{G} r_i$ and the standard deviation is defined as $\sigma = \sqrt{\frac{1}{G}\sum_{i=1}^{G}(r_i - \mu)^2}$. The default normalized advantage for the i^{th} rollout is defined as the following formulation:

$$\hat{A}_i = \frac{r_i - \mu}{\sigma}. \tag{2}$$

3.2 Proposed Method

Observation. We model the trajectory outcome as a Bernoulli random variable, $X \sim \text{Bernoulli}(p)$, $X \in \{0, 1\}$, where $X = 1$ denotes a successful trajectory and $X = 0$ denotes a failure. The success probability p is defined by the expectation of X, $\mathbb{E}[X] = p$, and the variance of this distribution is $\text{Var}(X) = p(1-p)$, which quantifies the certainty of the trajectory outcome, shown in fig. 1. However, directly measuring p is challenging, so we estimate it empirically using the ratio $p \approx \frac{N}{G}$, where G is the total number of sampled trajectories. N is the number of successful trajectories and is defined as the following formulation:

$$N = \sum_{i=1}^{G} \mathbf{1}_{\{r_i = 1\}}. \tag{3}$$

This empirical estimation approximates the true probability p via observed trajectories. Thus, we reveal that *samples exhibit varying certainty level within the GRPO sampling process*.

Advantage Percent Deviation. Sample would appear high certainty, when the prediction variance is close to zero ($\text{Var}(X) \to 0$), *i.e.*, $p \to 0$ or $p \to 1$, which typically corresponds to overly easy or difficult instances. In such cases, existing advantage faces two key challenges: *Advantage Reversion* and *Advantage Mirror*. Specifically, the advantage formulation, $\hat{A}_i = \frac{r_i - \mu}{\sigma}$, can produce misleading behaviors between trajectories with high and low certainty. For instance, in the case of *Advantage Reversion*, the high-certain trajectory with a relatively high reward of 0.9 in a batch of $\boldsymbol{r}_{\text{High}} = [0.9,\ 1,\ 1,\ 1]$ is assigned a large negative advantage ($\min \hat{A}_i = -1.73$), which is more extreme than the low-certain ones like $\boldsymbol{r}_{\text{Low}} = [0.1,\ 0.1,\ 1,\ 1]$, due to the small standard deviation exaggerating deviations from the mean. Similarly, as for *Advantage Mirror*, two reward batches that are symmetric around the center, such as $[0,\ 0.1,\ 0.1,\ 0.1]$ and $[0.9,\ 1,\ 1,\ 1]$, yield mirrored normalized advantage scores $[-1.73,\ 0.57,\ 0.57,\ 0.57]$, which makes semantically distinct cases to appear structurally equivalent normalization. These examples show how reliance on μ and σ alone can distort the relative evaluation of trajectories, especially when the variance is abnormally small or the rewards are symmetrically distributed, thus echoing a more robust advantage function.

Therefore, in our work, to address the question **i)**: *high-certainty samples advantage reconstruction*, we introduce the Advantage Percent Deviation (APD), to effectively address the issues of *Advantage Reversion* and *Advantage Mirror*. Instead of relying on z-score normalization, APD measures the relative deviation of each trajectory reward from the batch mean reward, formulated as follows:

$$\hat{A}_i^{APD} = \frac{r_i - \mu}{\mu}. \tag{4}$$

Methods	Advantage Formulation	Advantage Problem	
		Adv. Reversion	Adv. Mirror
GRPO ♣	$\hat{A}_i = \frac{r_i - \mu}{\sigma}$	✓	✓
Dr. GRPO ♦	$\hat{A}_i = r_i - \mu$		✓
GPG ♥	$\hat{A}_i = \alpha * (r_i - \mu)$		✓
TreeRPO ♠	$\hat{A}_i = \frac{r_i - \mu}{\mu(1-\mu)}$		✓
MAPO ★	$\hat{A}_i = (1-\lambda(p))\frac{r_i-\mu}{\sigma} + \lambda(p)\frac{r_i-\mu}{\mu}$	-	-

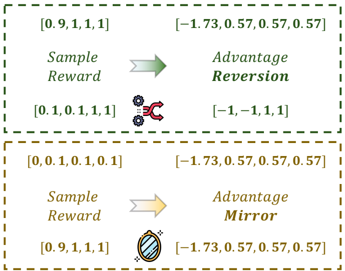


Figure 3: **Discussion on existed advantage functions**. $\boldsymbol{r}$ means the group reward. We denote $\mu = mean(\boldsymbol{r})$ and $\sigma = std(\boldsymbol{r})$. The reward is defined as a combination of the format reward (r_{Format}) and the accuracy reward ($r_{Accuracy}$), with a weighting factor of $\beta = 0.9$, *i.e.*, $r = (1-\beta)r_{Format} + \beta r_{Accuracy}$. $\alpha = 0.6$ is the valid sample rescale parameter for GPG Chu et al. (2025b). Refer to Sec. 3.3 for details.

This design emphasizes the proportional difference between individual rewards and the central tendency, ensuring that the advantage reflects not only the relative ordering but also the magnitude of deviation in percentage terms. By doing so, APD mitigates the instability caused by abnormally small standard deviations and prevents mirrored advantage allocation from being treated as equivalent, thereby providing a more stable and reasonable trajectory quality evaluation.

Trajectory Certainty Reweight. Considering that samples are in various trajectory certainty conditions, it is essential to dynamically adjust the advantage formulation across different samples. We propose the Trajectory Certainty Reweight (TCR), which adaptively reconstructs the advantage function based on trajectory certainty to address **ii)** *various-certainty sample advantage reweighting*. This design ensures that sample-specific characteristics are preserved, leading to a more faithful and stable evaluation of trajectory quality.

To be precise, we formalize TCR by introducing a certainty-aware weighting scheme. The key intuition is that when a trajectory exhibits high uncertainty (immature stage), the advantage should rely more on a variance-sensitive formulation (eq. (2)), while in highly certain (mature) stages it should instead emphasize a mean-relative formulation (eq. (4)) that remains stable even when variance collapses. To operationalize this idea, we use the estimated trajectory certainty p to interpolate these two advantages for different samples. We denote the trajectory certainty degree as follows:

$$\lambda(p) = 1 - 4p(1-p) \in [0,1] \quad (p \in [0,1]). \tag{5}$$

And, then we further construct the sample-wise advantage construction as follows:

$$\hat{A}_i^* = (1-\lambda(p)) * \underbrace{\boxed{\frac{r_i - \mu}{\sigma}}}_{\textbf{Deviation-based}} + \lambda(p) * \underbrace{\boxed{\frac{r_i - \mu}{\mu}}}_{\textbf{Mean-based}}. \tag{6}$$

The standard deviation–based advantage is weighted by $1 - \lambda(p)$, while the complementary factor $\lambda(p)$ is assigned to the mean-based advantage. In this way, the contribution shifts smoothly from deviation-based signals under uncertainty to mean-based signals under certainty, ensuring a balanced and robust construction of the advantage function across different trajectory certainty levels. Thus, we replace the original advantage $\hat{A}_i$ to proposed $\hat{A}_i^*$ in eq. (6) in eq. (1) for optimization. As a result, our method reveals the trajectory certainty phenomenon and effectively mitigates existing advantage limitations via dynamical advantage reweight operation. We provide the methodological framework in fig. 2 and the algorithm description in algorithm 1.

3.3 Discussion and Limitation

Advantage Exploration. The advantage function typically relies on group-based estimation from trajectory rewards $r_i \in \boldsymbol{r}$. Existing explorations Liu et al. (2025e); Chu et al. (2025b); Yang et al. (2025c) can be broadly classified as the following directions. ***First***, methods such as Dr. GRPO Liu et al. (2025e), GPG Chu et al. (2025b), S-GRPO Dai et al. (2025b), and wd1 Tang et al. (2025) remove the standard variance normalization term to alleviate reward bias. More recently, Yang et al. (2025c) identifies that conventional normalization fails to scale advantages properly under continuous rewards, and proposes to rewrite the variance as $\sigma = \mu(1-\mu)$ to address this issue. Meanwhile, KRPO Wang et al. (2025a) introduces a lightweight Kalman filter to dynamically estimate latent reward mean and variance, enabling more adaptive advantage normalization. However, these approaches fail to resolve both *advantage reversion* and *advantage mirror* problems in fig. 3.

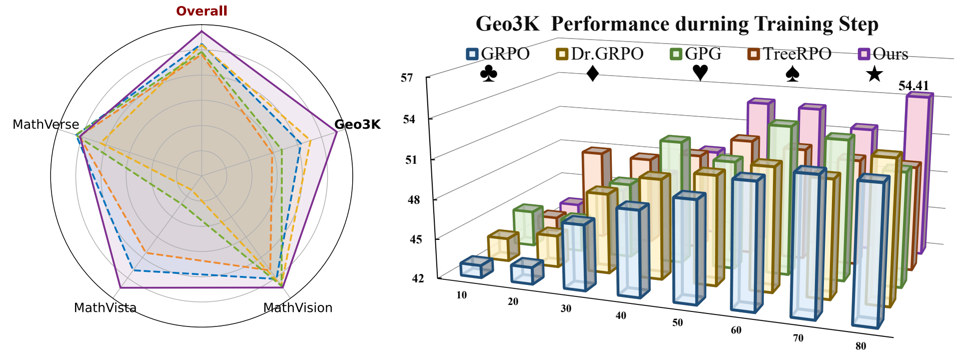


Figure 4: **Performance comparison with different advantage formulations** on the geometry task based on Qwen2.5-VL-7B-Instruct model with rollout number $G = 12$. Please refer to Sec. 3.3 for details.

Moreover, existing methods typically employ a uniform advantage formulation across all samples, overlooking the uniqueness of individual sample conditions. In contrast, we observe that samples exhibit varying degrees of certainty during optimization. Motivated by this insight, we propose a novel advantage function that leverages relative deviation for high-certainty samples, and further introduce a certainty-aware reweighting scheme that dynamically adjusts the advantage construction based on trajectory certainty. This design ensures a more faithful and stable evaluation of the sample situation across diverse training conditions. We further conduct the empirical experiments to validate the proposed mixture advantage solutions in fig. 4 and MAPO achieves a satisfying performance.

Conceptual Difference. Utilizing GRPO to enhance foundation models with reasoning capabilities has gained significant attention Liu et al. (2025e); Chen et al. (2025b); Chu et al. (2025a); Ma et al. (2025). Recent studies have explored the contribution of sample selection to the effectiveness of GRPO across three main streams. First, one line focuses on highlighting relatively simple samples to achieve a stable optimization. For instance, SEED-GRPO Chen et al. (2025a) utilizes semantic entropy to measure answer diversity and applies more conservative updates to hard questions. However, blindly emphasizing easy samples can restrict the model exploration ability Yue et al. (2025) and face model entropy collapse Zhang et al. (2024a). Second, another paradigm seeks to highlight hard samples. For example, GRPO-LEAD amplifies learning signals for challenging problems using a difficulty-aware advantage reweight. Recently, Pikus et al. (2025) has pointed out that the hardest examples consistently yield superior performance on reasoning benchmarks. However, this approach leads to longer convergence times. The third group focuses on eliminating meaningless samples. DAPO Yu et al. (2025) and GPG Chu et al. (2025b) aim to discard samples with vanishing advantages, *i.e.*, $\sigma = 0$. DAPO considers over-sampling and filtering out prompts with mean accuracy $\mu \in \{0, 1\}$, but this operation is not efficient in terms of training time. This inefficiency arises because the time required to collect a batch of desired examples is uncontrollable and depends on the task difficulty. In contrast, GPG seeks more accurate gradient estimation by rescaling the gradient based on the valid samples ratio, with a validity threshold of 0.6. In summary, existing work conducts a **monotonic emphasis** based on sample difficulty, which inevitably faces the prisoner dilemma of sample difficulty. In contrast, our work considers trajectory certainty and allocates different mixture ratios for high- and low-certainty samples, thereby introducing a **discriminative emphasis**. We reveal the gradient of MAPO compared with GRPO. *Without loss of generality*, we simplify the gradient analysis by ignoring clipping and KL regularization and considering the reward as a Bernoulli variable. We define the ratio between the gradients of MAPO and GRPO as:

$$\varrho(p) \triangleq \frac{\nabla_\theta \mathcal{J}_{\text{MAPO}}}{\nabla_\theta \mathcal{J}_{\text{GRPO}}} = (1 - \lambda(p)) + \lambda(p)\sqrt{\frac{1-p}{p}}, \quad \lambda(p) = 1 - 4p(1-p). \tag{7}$$

By further analysis (see details in appendix C), we obtain the following formulation:

$$\begin{cases} \varrho(p) > 1, & p \in (0, \frac{1}{2}), \\ \varrho(p) = 1, & p = \frac{1}{2}, \\ 0 < \varrho(p) < 1, & p \in (\frac{1}{2}, 1). \end{cases} \tag{8}$$

This shows that the mixed reward, MAPO *implicitly assigns larger gradients to harder samples (with* $p < \frac{1}{2}$*) and smaller gradients to easier ones (with* $p > \frac{1}{2}$*)*, which aligns with prior insights that appropriately emphasizing difficult samples enhances the performance GRPO Pikus et al. (2025).

Table 2: **Ablative study of key modules** for MAPO viaQwen2.5-VL-7B-Instruct with rollout number $G = 12$. Incorporate **sole** Advantage Percent Deviation (APD) can be regarded as the advantage replacement. Involving **both** APD and TCR achieves a satisfying performance. Please refer to Sec. 4.2. for details.

APD	TCR	Geo3K	MathVision	MathVista	MathVerse	$\mathcal{A}^{\mathcal{T}}$	$\bar{\mathcal{A}}$	EmoSet	WEBEmo	Emotion6	$\mathcal{A}^{\mathcal{T}}$	$\bar{\mathcal{A}}$
Vanilla		-	-	-	-	-	-	-	-	-	-	-
		51.91	26.74	72.70	43.93	47.79	49.85	75.50	49.90	60.44	55.17	65.33
✓		50.92	26.61	73.00	44.59	48.07	49.49	77.20	50.55	61.95	56.25	66.72
✓	Rand	53.41	24.90	71.20	43.38	46.49	49.95	76.90	50.20	62.12	56.16	66.53
✓	✓	**54.41**	27.30	73.20	43.81	**48.10**	**51.26**	**77.86**	50.75	60.61	**55.68**	**66.77**

Limitation. Despite achieving satisfactory performance with free hyperparameters, our research has several limitations. First, our approach uses trajectory certainty to treat different samples selectively. In extreme reinforcement scenarios or when foundational model capabilities are limited, it becomes difficult to generate a diverse set of successful trajectories, as rollout may consistently fail. In such cases, our method could reduce to a single function strategy. Second, although our method assigns different reward mechanisms for samples with different trajectory maturity levels, a more refined reward allocation method is still worth exploring. Third, due to computational constraints, our experiments are limited to models with up to 7B parameters and datasets with a few thousand samples. Future work would aim to extend these findings to larger-scale scenarios.

4 Experiments

4.1 Experimental Setup

Environment and Datasets. We conduct experiments on two reasoning scenarios: mathematics and emotion. For training, we utilize Geo3K [arXiv'21] Lu et al. (2021) and EmoSet [ICCV'23] Yang et al. (2023). These two datasets are respectively comprised of $2.1K$ training samples. Furthermore, for the out-of-domain validation, we respectively adopt out-of-domain datasets in the math and emotion fields: MathVista [arXiv'23] Lu et al. (2023), MathVision [NeurIPS'24] Wang et al. (2024), MathVerse [ECCV'24] Zhang et al. (2024b), WEBEmo [ECCV'18] Panda et al. (2018), and Emotion6 [CVPR'15] Peng et al. (2015). We provide a detailed dataset illustration in appendix B.1.

Architecture and Counterparts. We utilize the popular open-source Qwen2.5-VL-7B-Instruct as the base (Vanilla) model, which exhibits strong foundational capabilities well-suited for subsequent RL training Yang et al. (2025b); Bai et al. (2025). We further conduct the comparison with the GRPO Shao et al. (2024) and DAPO Yu et al. (2025) to validate the effectiveness of our method.

Implementation Details. Experiments are conducted on 8 A100 GPUs. Detail is in appendix B.2.

Evaluation Metrics. We evaluate both in-domain ($\mathcal{A}^{\mathcal{S}}$) and out-of-domain ($\mathcal{A}^{\mathcal{T}}$). Let $\mathcal{T} = \{\mathcal{T}_t\}_{t=1}^{|\mathcal{T}|}$ represent the unseen dataset set and $\mathcal{S}$ denote the training distribution. Thus, we derive the following evaluation metrics forms $\mathcal{A}^{\mathcal{S}} = \text{Acc.}(\mathcal{S})$ and $\mathcal{A}^{\mathcal{T}} = \frac{1}{|\mathcal{T}|}\sum_t^{|\mathcal{T}|} \text{Acc.}(\mathcal{T}_i)$. Acc. denotes the accuracy metric. Furthermore, we use the Average metric to evaluate overall performance as $\bar{\mathcal{A}} = \frac{\mathcal{A}^{\mathcal{S}} + \mathcal{A}^{\mathcal{T}}}{2}$.

4.2 Diagnostic Analysis

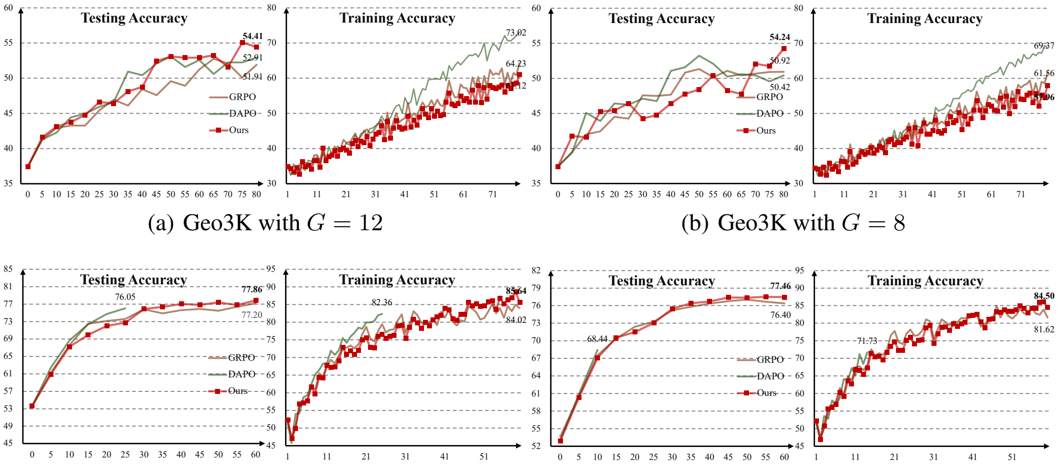
We perform ablation studies on the Geo3K and EmoSet datasets, utilizing the Qwen2.5-VL-7B-Instruct model to facilitate an in-depth analysis. We quantitatively analyze the proposed Mixed Advantage Policy Optimization (MAPO) in Tab. 2. The ablation demonstrates that solely replacing the advantage function from the original $\hat{A}_i = \frac{r_i - \mu}{\sigma}$ to $\hat{A}_i^{APD} = \frac{r_i - \mu}{\mu}$ leads to limited performance improvement or even degradation, which underscores the necessity for a dynamic advantage function. Furthermore, utilizing random weight allocation ($\lambda(p) \sim \mathcal{U}(0, 1)$ in eq. (6)) fails to achieve stable performance improvements. Thus, incorporating the Trajectory Certainty Reweight, which accounts for trajectory certainty, further enhances overall performance.

4.3 Comparison to State-of-the-Arts

We benchmark MAPO against state-of-the-art reinforcement learning frameworks in reasoning tasks. As illustrated in Tab. 3, MAPO consistently outperforms Vanilla, GRPO, and DAPO across

Table 3: **Performance comparison with GRPO variants** on the geometry and emotional reasoning tasks. We mark the **Best** in bold an <u>Second</u> in underline across different methods. Refer to Sec. 4.3.

Methods	Geo3K	MathVision	MathVista	MathVerse	$\mathcal{A}^{\mathcal{T}}$	$\bar{\mathcal{A}}$	EmoSet	WEBEmo	Emotion6	$\mathcal{A}^{\mathcal{T}}$	$\bar{\mathcal{A}}$
Vanilla	37.43	24.51	67.20	40.02	43.91	40.67	53.65	46.85	52.19	49.52	51.58
Qwen2.5-VL-7B-Instruct with Rollout Number $G = 12$											
GRPO	51.91	26.74	72.70	43.93	47.79	49.85	77.20	49.90	60.44	55.17	66.18
DAPO	<u>52.91</u>	26.51	73.50	44.59	**48.20**	<u>50.56</u>	76.05	50.60	<u>60.61</u>	<u>55.60</u>	65.82
MAPO	**54.41**	27.30	73.20	43.81	<u>48.10</u>	**51.26**	**77.86**	50.75	60.61	**55.68**	**66.77**
Qwen2.5-VL-7B-Instruct with Rollout Number $G = 8$											
GRPO	<u>50.92</u>	26.38	72.60	43.45	**47.48**	<u>49.20</u>	<u>76.40</u>	49.90	60.27	<u>55.08</u>	<u>65.74</u>
DAPO	50.42	26.41	72.40	43.15	47.32	48.87	68.44	47.80	58.08	52.94	60.69
MAPO	**54.24**	27.37	71.30	43.40	<u>47.36</u>	**50.80**	**77.46**	50.05	61.28	**55.66**	**66.56**



(a) Geo3K with $G = 12$

(b) Geo3K with $G = 8$

(c) EmoSet with $G = 12$

(d) EmoSet with $G = 8$

Figure 5: **Training and testing accuracy during the training process**. DAPO fails to conduct complete training due to dynamic sampling failure in EmoSet scenario. Refer to Sec. 4.3.

both in-domain, *e.g.*, Geo3K and out-of-domain, *e.g.*, MathVision, MathVista, and MathVerse, showing strong generalization performance under different rollout numbers. With $G = 12$, it achieves the highest overall accuracies (51.26 on math and 66.77 on emotion), and even with $G = 8$, it maintains superior results (50.80 and 66.55), validating that its mixed advantage formulation effectively mitigates advantage reversion and mirror issues while ensuring reliable optimization. Overall, it advances state-of-the-art performance with consistent gains across diverse reasoning tasks.

5 Conclusion

In our work, we focus on Group Relative Policy Optimization (GRPO) and observe that the advantage function plays a crucial role in evaluating trajectory importance. However, existing advantage formulations face two challenges: advantage reversion and advantage mirror. To address these issues, we propose Mixed Advantage Policy Optimization (MAPO). In particular, we uncover the trajectory certainty property and introduce advantage percent deviation for high-certainty trajectories. Furthermore, we dynamically reweight the advantage function according to trajectory certainty, thereby adaptively tailoring the advantage to sample-specific characteristics. Our method offers three key advantages: First, *No Architecture Dependency*: MAPO operates without additional model architectures, ensuring high transferability across different architectures. Second, *No Thinking Pattern Conflict*: our approach directly evaluates trajectory advantages while maintaining compatibility with diverse reasoning formats. Third, *No Hyper-Parameter Configuration*: by leveraging trajectory certainty to adaptively reweight sample advantage formulations, our method avoids the need for additional hyperparameters, thereby improving reinforcement effectiveness. MAPO has been validated across diverse scenarios, underscoring its potential for broader applications.

REPRODUCIBILITY

To facilitate the reproducibility, we provide the source code for both the training and evaluating framework in the supplementary material. All experimental settings, including key hyperparameter for training, are detailed in appendix B.2. We experiment on 8 NVIDIA A100 GPUs.

THE USE OF LARGE LANGUAGE MODELS (LLMS)

We disclose using LLMs solely to aid and polish writing—enhancing academic expression accuracy, argument coherence, and text logic. All core ideas, research designs, experiments, and conclusions are independently developed by the authors.

ETHICS AND SOCIAL IMPACT

This work is purely methodological, focusing on enhancing foundation model reasoning through Mixed Advantage Policy Optimization. It uses only public benchmark datasets and involves no personal data, or sensitive content. The method does not create or process private data, nor is it deployed in real-world applications. While stronger reasoning may indirectly influence downstream uses, our study does not explore deployment, bias, or misuse. The research is academic in purpose and poses no direct ethical or societal risks, aligning with responsible and trustworthy AI development.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, volume 35, pp. 23716–23736, 2022.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023a.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023b.

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.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, pp. 1877–1901, 2020.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pp. 9650–9660, 2021.

Changyu Chen, Zichen Liu, Chao Du, Tianyu Pang, Qian Liu, Arunesh Sinha, Pradeep Varakantham, and Min Lin. Bootstrapping language models with dpo implicit rewards. *arXiv preprint arXiv:2406.09760*, 2024.

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

Yi Chen, Yuying Ge, Rui Wang, Yixiao Ge, Junhao Cheng, Ying Shan, and Xihui Liu. Grpo-care: Consistency-aware reinforcement learning for multimodal reasoning. *arXiv preprint arXiv:2506.16141*, 2025b.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*, 2025a.

Xiangxiang Chu, Hailang Huang, Xiao Zhang, Fei Wei, and Yong Wang. Gpg: A simple and strong reinforcement learning baseline for model reasoning. *arXiv preprint arXiv:2504.02546*, 2025b.

Muzhi Dai, Shixuan Liu, and Qingyi Si. Stable reinforcement learning for efficient reasoning. *arXiv preprint arXiv:2505.18086*, 2025a.

Muzhi Dai, Chenxu Yang, and Qingyi Si. S-grpo: Early exit via reinforcement learning in reasoning models. *arXiv preprint arXiv:2505.07686*, 2025b.

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

Fei Ding, Baiqiao Wang, Zijian Zeng, and Youwei Wang. Multi-layer grpo: Enhancing reasoning and self-correction in large language models. *arXiv preprint arXiv:2506.04746*, 2025.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

Kaixuan Fan, Kaituo Feng, Haoming Lyu, Dongzhan Zhou, and Xiangyu Yue. Sophiavl-r1: Reinforcing mllms reasoning with thinking reward. *arXiv preprint arXiv:2505.17018*, 2025.

Yunhao Fang, Ligeng Zhu, Yao Lu, Yan Wang, Pavlo Molchanov, Jang Hyun Cho, Marco Pavone, Song Han, and Hongxu Yin. $vila^2$: Vila augmented vila. *arXiv preprint arXiv:2407.17453*, 2024.

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

Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. *arXiv preprint arXiv:2503.24290*, 2025.

Qihan Huang, Weilong Dai, Jinlong Liu, Wanggui He, Hao Jiang, Mingli Song, Jingyuan Chen, Chang Yao, and Jie Song. Boosting mllm reasoning with text-debiased hint-grpo. In *ICCV*, 2025.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pp. 12888–12900. PMLR, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, pp. 19730–19742. PMLR, 2023.

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. *arXiv preprint arXiv:2506.09736*, 2025.

Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *CVPR*, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023b.

Shunyu Liu, Wenkai Fang, Zetian Hu, Junjie Zhang, Yang Zhou, Kongcheng Zhang, Rongcheng Tu, Ting-En Lin, Fei Huang, Mingli Song, et al. A survey of direct preference optimization. *arXiv preprint arXiv:2503.11701*, 2025a.

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. *arXiv preprint arXiv:2504.13055*, 2025b.

Yuqi Liu, Tianyuan Qu, Zhisheng Zhong, Bohao Peng, Shu Liu, Bei Yu, and Jiaya Jia. Vision-reasoner: Unified visual perception and reasoning via reinforcement learning. *arXiv preprint arXiv:2505.12081*, 2025c.

Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, et al. Nvila: Efficient frontier visual language models. In *CVPR*, 2025d.

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. *arXiv preprint arXiv:2503.20783*, 2025e.

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

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. *arXiv preprint arXiv:2105.04165*, 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. *arXiv preprint arXiv:2310.02255*, 2023.

Lu Ma, Hao Liang, Meiyi Qiang, Lexiang Tang, Xiaochen Ma, Zhen Hao Wong, Junbo Niu, Chengyu Shen, Runming He, Bin Cui, et al. Learning what reinforcement learning can't: Interleaved online fine-tuning for hardest questions. *arXiv preprint arXiv:2506.07527*, 2025.

OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

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

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

Benjamin Pikus, Pratyush Ranjan Tiwari, and Burton Ye. Hard examples are all you need: Maximizing grpo post-training under annotation budgets. *arXiv preprint arXiv:2508.14094*, 2025.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pp. 8748–8763, 2021.

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

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

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.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *EuroSys*, pp. 1279–1297, 2025.

Xiaohang Tang, Rares Dolga, Sangwoong Yoon, and Ilija Bogunovic. wd1: Weighted policy optimization for reasoning in diffusion language models. *arXiv preprint arXiv:2507.08838*, 2025.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025.

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

Hu Wang, Congbo Ma, Ian Reid, and Mohammad Yaqub. Kalman filter enhanced grpo for reinforcement learning-based language model reasoning. *arXiv preprint arXiv:2505.07527*, 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*, pp. 95095–95169, 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. *arXiv preprint arXiv:2508.18265*, 2025b.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a foreign language: BEiT pretraining for vision and vision-language tasks. In *CVPR*, 2023.

Liang Wen, Yunke Cai, Fenrui Xiao, Xin He, Qi An, Zhenyu Duan, Yimin Du, Junchen Liu, Lifu Tang, Xiaowei Lv, et al. Light-r1: Curriculum sft, dpo and rl for long cot from scratch and beyond. *arXiv preprint arXiv:2503.10460*, 2025.

Wei Xiong, Jiarui Yao, Yuhui Xu, Bo Pang, Lei Wang, Doyen Sahoo, Junnan Li, Nan Jiang, Tong Zhang, Caiming Xiong, et al. A minimalist approach to llm reasoning: from rejection sampling to reinforce. *arXiv preprint arXiv:2504.11343*, 2025.

Zhongwen Xu and Zihan Ding. Single-stream policy optimization. *arXiv preprint arXiv:2509.13232*, 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. *arXiv preprint arXiv:2505.09388*, 2025a.

An Yang, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoyan Huang, Jiandong Jiang, Jianhong Tu, Jianwei Zhang, Jingren Zhou, et al. Qwen2. 5-1m technical report. *arXiv preprint arXiv:2501.15383*, 2025b.

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

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

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

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

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837*, 2025.

Chuheng Zhang, Wei Shen, Li Zhao, Xuyun Zhang, Xiaolong Xu, Wanchun Dou, and Jiang Bian. Policy filtration for rlhf to mitigate noise in reward models. *arXiv preprint arXiv:2409.06957*, 2024a.

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

Jixiao Zhang and Chunsheng Zuo. Grpo-lead: A difficulty-aware reinforcement learning approach for concise mathematical reasoning in language models. *arXiv preprint arXiv:2504.09696*, 2025.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *ECCV*, pp. 169–186. Springer, 2024b.

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

Yichen Zhu, Minjie Zhu, Ning Liu, Zhicai Ou, Xiaofeng Mou, and Jian Tang. Llava-phi: Efficient multi-modal assistant with small language model. *arXiv preprint arXiv:2401.02330*, 2024.

Appendix

A Notation and Algorithm

We provide the notation table in Tab. 4 and proposed method algorithm in algorithm 1.

Algorithm 1: MAPO

```
Input: Reference model $\pi_{ref}$, old model $\pi_{old}$, current policy model $\pi_\theta$, group size $G$, Training Step $E$,
       Current Step $e$, Training Batch $B$, Question Distribution $\rho_Q$, Query Prompt $q$
Initialize $\pi_\theta \leftarrow \pi_{ref}$
for $e = 1, 2, ..., E$ do
    $\pi_{old} \leftarrow \pi_\theta$
    $q \sim \rho_Q, o = \{o_i\}_{i=1}^{G} \sim \pi_{old}(\cdot|q)$ ;                // Sample prompt with $G$ trajectory.
    $\boldsymbol{r} = \{r_i\}_{i=1}^{G} = R(o)$ ;                                     // Measure trajectory reward.
    $\mu = \frac{1}{G}\sum_{r_i \in \boldsymbol{r}} r_i, \sigma = \sqrt{\frac{1}{G}\sum_{i=1}^{G}(r_i - \mu)^2}$ ;   // Calculate static information.
    Advantage Percent Deviation;
    $\hat{A}_i \leftarrow (\boldsymbol{r}, \mu, \sigma)$ via eq. (2), $\hat{A}_i^{APD} \leftarrow (\boldsymbol{r}, \mu)$ via eq. (4) ;   // Measure advantage.
    Trajectory Certainty Reweight;
    $N = \sum_{i=1}^{G} \mathbf{1}_{\{r_i=1\}}, p = \frac{N}{G}$ ;                   // Calculate trajectory success ratio.
    $\lambda(p) = 1 - 4p(1-p)$ ;                                                      // Measure trajectory maturity degree.
    $\hat{A}_i^* \leftarrow (\hat{A}_i, \hat{A}_i^{APD}, \lambda)$ through eq. (6) ;  // Mixed advantage.
    $\mathcal{J}_{GRPO}(\theta) \leftarrow (\hat{A}_i^*, o, \pi_\theta, \pi_{old}, \pi_{ref}), \theta = \theta - \eta\nabla\mathcal{J}_{GRPO}(\theta)$ ;   // Update Weight.
end
```

Table 4: **Notation used in MAPO**. Summary of key variables and operations in our method. The *Definition* column indicates where each symbol first appears in the main text.

Symbol	Description	Definition
q	Query prompt sampled from distribution ρ_Q	eq. (1)
o_i	i-th trajectory (rollout) sampled from π_{old}	eq. (1)
r_i	Reward assigned to trajectory o_i	eq. (1)
G	Group size (number of rollouts per query)	eq. (1)
$\pi_\theta, \pi_{\text{old}}, \pi_{\text{ref}}$	Current, old, and reference policy models	eq. (1)
$J_{\text{GRPO}}(\theta)$	Group Relative Policy Optimization objective	eq. (1)
β	KL regularization coefficient	eq. (1)
$f_\epsilon(x, y)$	Clipping function $\min(xy, \text{clip}(x, 1-\epsilon, 1+\epsilon)y)$	eq. (1)
$\mathbb{D}_{KL}[\pi_\theta \| \pi_{\text{ref}}]$	KL divergence between policy and reference model	eq. (1)
$R(q, o_i)$	Reward function	Sec. 3.1
μ, σ	Mean and standard deviation of rewards in group	eq. (2)
$\hat{A}_i$	Standardized advantage $\frac{r_i - \mu}{\sigma}$	eq. (2)
N	Number of successful trajectories in a group	eq. (3)
p	Empirical success ratio $p = \frac{N}{G}$	eq. (3)
$\hat{A}_i^{\text{APD}}$	Advantage Percent Deviation $\frac{r_i - \mu}{\mu + \epsilon}$	eq. (4)
$\lambda(p)$	Trajectory maturity degree $1 - 4p(1-p)$	eq. (5)
$\hat{A}_i^*$	Mixed advantage combining $\hat{A}_i$ and $\hat{A}_i^{\text{APD}}$	eq. (6)
$r_{\text{Format}}, r_{\text{Accuracy}}$	Format reward and accuracy reward	fig. 3
$\varrho(p)$	Gradient ratio $\nabla_\theta J_{\text{MAPO}} / \nabla_\theta J_{\text{GRPO}}$	eq. (7)
$\mathcal{S}$	Training distribution	Sec. 4.1
$\mathcal{T} = \{T_t\}_{t=1}^{\|\mathcal{T}\|}$	Set of unseen test datasets	Sec. 4.1
$\|\mathcal{T}\|$	Number of unseen test datasets	Sec. 4.1
$\mathcal{A}^S, \mathcal{A}^T, \bar{\mathcal{A}}$	In-domain, out-of-domain and average accuracy	Sec. 4.1

B Experimental Informataion

B.1 Dataset Introduction

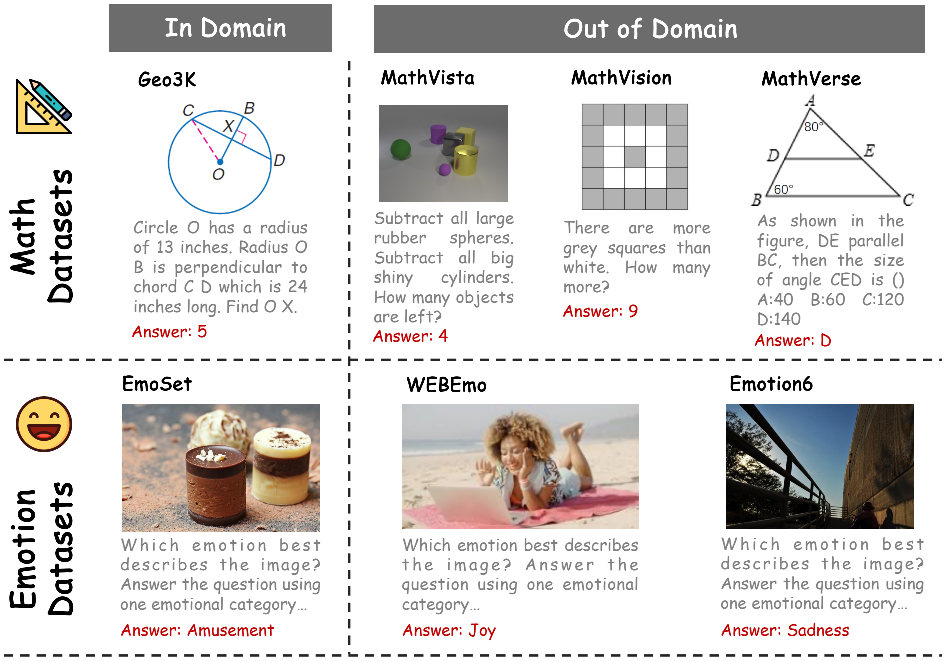
We use the following two datasets from the mathematics and emotional tasks for experiments.

- 📐 Geo3K [arXiv'21] Lu et al. (2021) Designed for geometry problem solving, this dataset contains images, text, and formulas that require models to perform joint visual–symbolic reasoning.
- 😄 EmoSet [ICCV'23] Yang et al. (2023) This large-scale collection targets visual emotion recognition, covering diverse scenes and a broad range of emotion categories.

Furthermore, for above two scenarios, we respectively conduct the evaluation on the following out-of-domain datasets to vaildate its generalization ability.

- 📐 MathVista [arXiv'23] Lu et al. (2023) Serving as a benchmark for visual mathematical reasoning, spanning algebra, geometry, and word problems for cross-domain generalization.
- 📐 MathVision [NeurIPS'24] Wang et al. (2024) Proposed for multimodal mathematical reasoning, it emphasizes inference across visual diagrams and natural language expressions.
- 📐 MathVerse [ECCV'24] Zhang et al. (2024b) Built to assess model understanding of complex charts, geometric figures, and formula-rich inputs, emphasizing visual interpretation in reasoning.
- 😄 WEBEmo [ECCV'18] Panda et al. (2018) Comprising millions of web images, this dataset spans 7 high-level emotion categories and supports recognition and cross-domain emotion analysis.
- 😄 Emotion6 [CVPR'15] Peng et al. (2015) A classic benchmark for visual emotion recognition, consisting of 6 basic emotion categories and widely used for standard evaluation.

We plot a detailed dataset case illustration in fig. 6.



Circle O has a radius of 13 inches. Radius O B is perpendicular to chord C D which is 24 inches long. Find O X.
Answer: 5

Subtract all large rubber spheres. Subtract all big shiny cylinders. How many objects are left?
Answer: 4

There are more grey squares than white. How many more?
Answer: 9

As shown in the figure, DE parallel BC, then the size of angle CED is ()
A:40 B:60 C:120 D:140
Answer: D

Which emotion best describes the image? Answer the question using one emotional category...
Answer: Amusement

Which emotion best describes the image? Answer the question using one emotional category...
Answer: Joy

Which emotion best describes the image? Answer the question using one emotional category...
Answer: Sadness

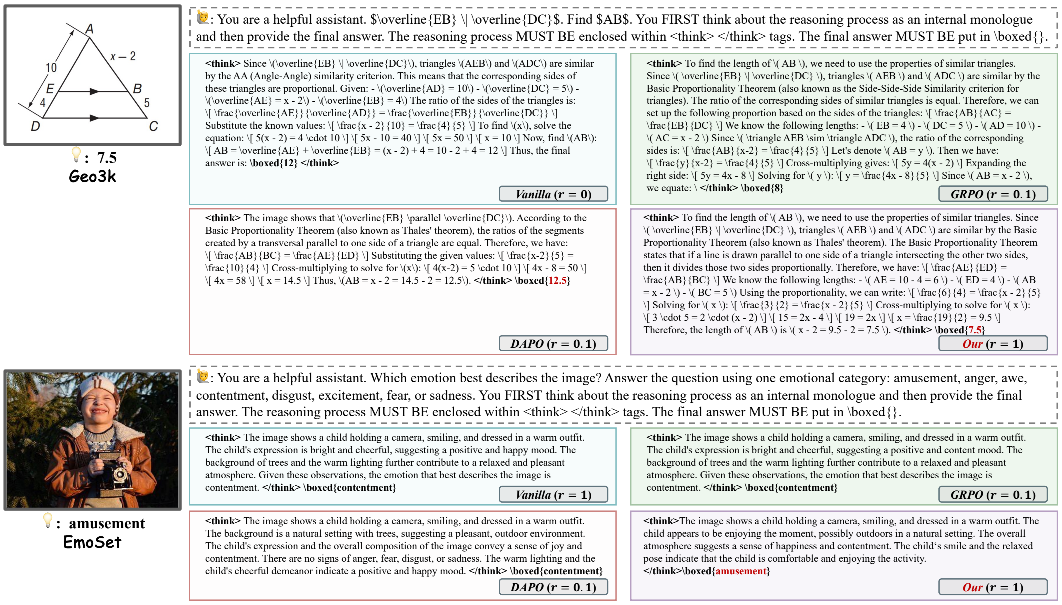
Figure 6: **Data Visualization** from the In Domain and Out of Domain datasets used in our experiments.

B.2 Implementation Details

We conduct the experiments on the EasyR1[1] Zheng et al. (2025) as our reinforcement learning training framework, which is built on Verl[2] Sheng et al. (2025). The rollout batch size is set to 512, and the global batch size is 128. The rollout temperature during training is fixed at 1.0, with Top-p set to 0.99. To mitigate token-length bias, we compute the policy loss using a token-mean aggregation strategy. The vision tower of Qwen2.5-VL-7B is fine-tuned without freezing. The optimizer is AdamW Kingma & Ba (2014) with a learning rate of 1×10^{-6}, and the KL coefficient β is set to 1×10^{-2}. For the validation setting, we set the temperature to 0.5. The maximum number of tokens to generate is 2048, and Top-p sampling is 0.95. The training epoch is respectively set as $T = 20$ and $T = 15$ for Geo3K and EmoSet. Consequently, the training step is $E = 80$ for Geo3K and $E = 60$ EmoSet. These configurations are consistent with the EasyR1.

[1] https://github.com/hiyouga/EasyR1

[2] https://github.com/volcengine/verl



: You are a helpful assistant. $\overline{EB} \| \overline{DC}$. Find $AB$. 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{}.

<think> Since \(\overline{EB} \| \overline{DC}\), triangles \(AEB\) and \(ADC\) are similar by the AA (Angle-Angle) similarity criterion. This means that the corresponding sides of these triangles are proportional. Given: - \(\overline{AD} = 10\) - \(\overline{DC} = 5\) - \(\overline{AE} = x - 2\) - \(\overline{EB} = 4\) The ratio of the sides of the triangles is: \[\frac{\overline{AE}}{\overline{AD}} = \frac{\overline{EB}}{\overline{DC}} \] Substitute the known values: \[\frac{x - 2}{10} = \frac{4}{5} \] To find \(x\), solve the equation: \[5(x - 2) = 4 \cdot 10 \] \[5x - 10 = 40 \] \[5x = 50 \] \[x = 10 \] Now, find \(AB\): \[AB = \overline{AE} + \overline{EB} = (x - 2) + 4 = 10 - 2 + 4 = 12 \] Thus, the final answer is: **\boxed{12} </think>**

***Vanilla* (r = 0)**

<think> To find the length of \(AB \), we need to use the properties of similar triangles. Since \(\overline{EB} \| \overline{DC} \), triangles \(AEB \) and \(ADC \) are similar by the Basic Proportionality Theorem (also known as the Side-Side-Side Similarity criterion for triangles). The ratio of the corresponding sides of similar triangles is equal. Therefore, we can set up the following proportion based on the sides of the triangles: \[\frac{AB}{AC} = \frac{EB}{DC} \] We know the following lengths: - \(EB = 4 \) - \(DC = 5 \) - \(AD = 10 \) - \(AC = x - 2 \) Since \(\triangle AEB \sim \triangle ADC \), the ratio of the corresponding sides is: \[\frac{AB}{x-2} = \frac{4}{5} \] Let's denote \(AB = y \). Then we have: \[\frac{y}{x-2} = \frac{4}{5} \] Cross-multiplying gives: \[5y = 4(x - 2) \] Expanding the right side: \[5y = 4x - 8 \] Solving for \(y \): \[y = \frac{4x - 8}{5} \] Since \(AB = x - 2 \), we equate: \ **</think> \boxed{8}**

***GRPO* (r = 0.1)**

<think> The image shows that \(\overline{EB} \parallel \overline{DC}\). According to the Basic Proportionality Theorem (also known as Thales' theorem), the ratios of the segments created by a transversal parallel to one side of a triangle are equal. Therefore, we have: \[\frac{AB}{BC} = \frac{AE}{ED} \] Substituting the given values: \[\frac{x-2}{5} = \frac{10}{4} \] Cross-multiplying to solve for \(x\): \[4(x-2) = 5 \cdot 10 \] \[4x - 8 = 50 \] \[4x = 58 \] \[x = 14.5 \] Thus, \(AB = x - 2 = 14.5 - 2 = 12.5\). **</think> \boxed{12.5}**

***DAPO* (r = 0.1)**

<think> To find the length of \(AB \), we need to use the properties of similar triangles. Since \(\overline{EB} \| \overline{DC} \), triangles \(AEB \) and \(ADC \) are similar by the Basic Proportionality Theorem (also known as Thales' theorem). The Basic Proportionality Theorem states that if a line is drawn parallel to one side of a triangle intersecting the other two sides, then it divides those two sides proportionally. Therefore, we have: \[\frac{AE}{ED} = \frac{AB}{BC} \] We know the following lengths: - \(AE = 10 - 4 = 6 \) - \(ED = 4 \) - \(AB = x - 2 \) - \(BC = 5 \) Using the proportionality, we can write: \[\frac{6}{4} = \frac{x - 2}{5} \] Solving for \(x \): \[\frac{3}{2} = \frac{x - 2}{5} \] Cross-multiplying to solve for \(x \): \[3 \cdot 5 = 2 \cdot (x - 2) \] \[15 = 2x - 4 \] \[19 = 2x \] \[x = \frac{19}{2} = 9.5 \] Therefore, the length of \(AB \) is \(x - 2 = 9.5 - 2 = 7.5 \). **</think> \boxed{7.5}**

***Our* (r = 1)**

: You are a helpful assistant. Which emotion best describes the image? Answer the question using one emotional category: amusement, anger, awe, contentment, disgust, excitement, fear, or sadness. 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{}.

<think> The image shows a child holding a camera, smiling, and dressed in a warm outfit. The child's expression is bright and cheerful, suggesting a positive and happy mood. The background of trees and the warm lighting further contribute to a relaxed and pleasant atmosphere. Given these observations, the emotion that best describes the image is contentment. **</think> \boxed{contentment}**

***Vanilla* (r = 1)**

<think> The image shows a child holding a camera, smiling, and dressed in a warm outfit. The child's expression is bright and cheerful, suggesting a positive and content mood. The background of trees and the warm lighting further contribute to a relaxed and pleasant atmosphere. Given these observations, the emotion that best describes the image is contentment. **</think> \boxed{contentment}**

***GRPO* (r = 0.1)**

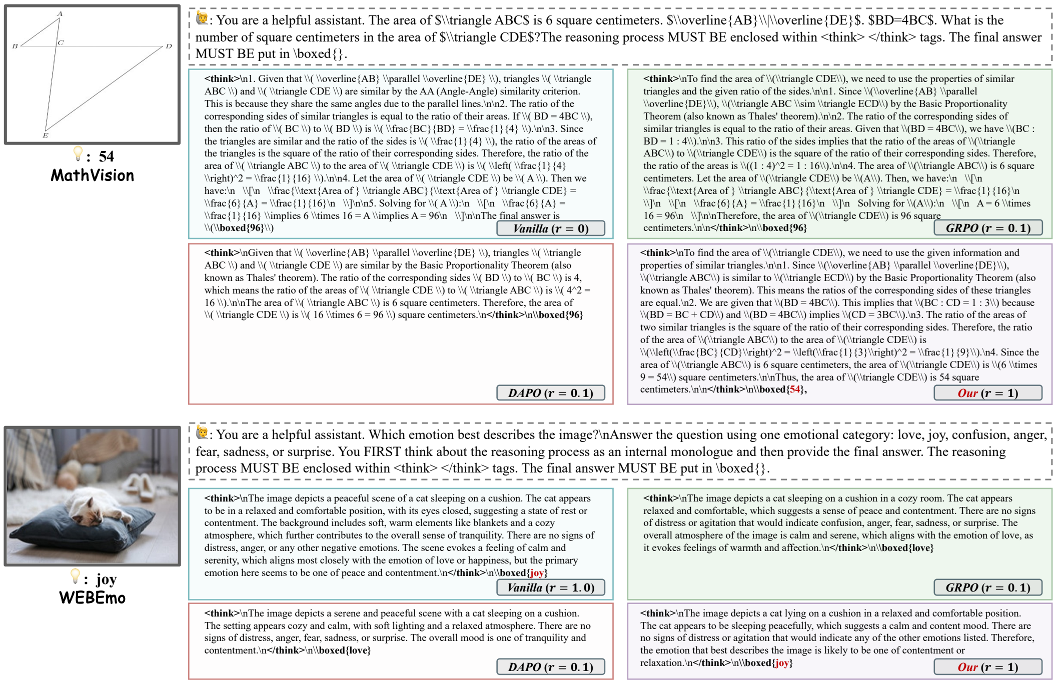
<think> The image shows a child holding a camera, smiling, and dressed in a warm outfit. The background is a natural setting with trees, suggesting a pleasant, outdoor environment. The child's expression and the overall composition of the image convey a sense of joy and contentment. There are no signs of anger, fear, disgust, or sadness. The warm lighting and the child's cheerful demeanor indicate a positive and happy mood. **</think> \boxed{contentment}**

***DAPO* (r = 0.1)**

<think>The image shows a child holding a camera, smiling, and dressed in a warm outfit. The child appears to be enjoying the moment, possibly outdoors in a natural setting. The overall atmosphere suggests a sense of happiness and contentment. The child's smile and the relaxed pose indicate that the child is comfortable and enjoying the activity. **</think>\boxed{amusement}**

***Our* (r = 1)**

Figure 7: **In-Domain Case visualization**.



: You are a helpful assistant. The area of $\\triangle ABC$ is 6 square centimeters. $\\overline{AB}\\|\\overline{DE}$. $BD=4BC$. What is the number of square centimeters in the area of $\\triangle CDE$?The reasoning process MUST BE enclosed within <think> </think> tags. The final answer MUST BE put in \boxed{}.

<think>\n1. Given that \\(\\overline{AB} \\parallel \\overline{DE} \\), triangles \\(\\triangle ABC \\) and \\(\\triangle CDE \\) are similar by the AA (Angle-Angle) similarity criterion. This is because they share the same angles due to the parallel lines.\n\n2. The ratio of the corresponding sides of similar triangles is equal to the ratio of their areas. If \\(BD = 4BC \\), then the ratio of \\(BC \\) to \\(BD \\) is \\(\\frac{BC}{BD} = \\frac{1}{4} \\).\n\n3. Since the triangles are similar and the ratio of the sides is \\(\\frac{1}{4} \\), the ratio of the areas of the triangles is the square of the ratio of their corresponding sides. Therefore, the ratio of the area of \\(\\triangle ABC \\) to the area of \\(\\triangle CDE \\) is \\(\\left(\\frac{1}{4} \\right)^2 = \\frac{1}{16} \\).\n\n4. Let the area of \\(\\triangle CDE \\) be \\(A \\). Then we have:\n   \\[\n   \\frac{\\text{Area of } \\triangle ABC}{\\text{Area of } \\triangle CDE} = \\frac{6}{A} = \\frac{1}{16}\n   \\]\n\n5. Solving for \\(A \\):\n   \\[\n   \\frac{6}{A} = \\frac{1}{16} \\implies 6 \\times 16 = A \\implies A = 96\n   \\]\n\nThe final answer is \\(**\\boxed{96}**\\)

***Vanilla* (r = 0)**

<think>\nTo find the area of \\(\\triangle CDE\\), we need to use the properties of similar triangles and the given ratio of the sides.\n\n1. Since \\(\\overline{AB} \\parallel \\overline{DE}\\), \\(\\triangle ABC \\sim \\triangle ECD\\) by the Basic Proportionality Theorem (also known as Thales' theorem).\n\n2. The ratio of the corresponding sides of similar triangles is equal to the ratio of their areas. Given that \\(BD = 4BC\\), we have \\(BC : BD = 1 : 4\\).\n\n3. This ratio of the sides implies that the ratio of the areas of \\(\\triangle ABC\\) to \\(\\triangle CDE\\) is the square of the ratio of their corresponding sides. Therefore, the ratio of the areas is \\((1 : 4)^2 = 1 : 16\\).\n\n4. The area of \\(\\triangle ABC\\) is 6 square centimeters. Let the area of \\(\\triangle CDE\\) be \\(A\\). Then, we have:\n   \\[\n   \\frac{\\text{Area of } \\triangle ABC}{\\text{Area of } \\triangle CDE} = \\frac{1}{16}\n   \\]\n   \\[\n   \\frac{6}{A} = \\frac{1}{16}\n   \\]\n   Solving for \\(A\\):\n   \\[\n   A = 6 \\times 16 = 96\n   \\]\n\nTherefore, the area of \\(\\triangle CDE\\) is 96 square centimeters.\n\n**</think>**\n\\boxed{96}

***GRPO* (r = 0.1)**

<think>\nGiven that \\(\\overline{AB} \\parallel \\overline{DE} \\), triangles \\(\\triangle ABC \\) and \\(\\triangle CDE \\) are similar by the Basic Proportionality Theorem (also known as Thales' theorem). The ratio of the corresponding sides \\(BD \\) to \\(BC \\) is 4, which means the ratio of the areas of \\(\\triangle CDE \\) to \\(\\triangle ABC \\) is \\(4^2 = 16 \\).\n\nThe area of \\(\\triangle ABC \\) is 6 square centimeters. Therefore, the area of \\(\\triangle CDE \\) is \\(16 \\times 6 = 96 \\) square centimeters.\n**</think>**\n\\boxed{96}

***DAPO* (r = 0.1)**

<think>\nTo find the area of \\(\\triangle CDE\\), we need to use the given information and properties of similar triangles.\n\n1. Since \\(\\overline{AB} \\parallel \\overline{DE}\\), \\(\\triangle ABC\\) is similar to \\(\\triangle ECD\\) by the Basic Proportionality Theorem (also known as Thales' theorem). This means the ratios of the corresponding sides of these triangles are equal.\n2. We are given that \\(BD = 4BC\\). This implies that \\(BC : CD = 1 : 3\\) because \\(BD = BC + CD\\) and \\(BD = 4BC\\) implies \\(CD = 3BC\\).\n3. The ratio of the areas of two similar triangles is the square of the ratio of their corresponding sides. Therefore, the ratio of the area of \\(\\triangle ABC\\) to the area of \\(\\triangle CDE\\) is \\(\\left(\\frac{BC}{CD}\\right)^2 = \\left(\\frac{1}{3}\\right)^2 = \\frac{1}{9}\\).\n4. Since the area of \\(\\triangle ABC\\) is 6 square centimeters, the area of \\(\\triangle CDE\\) is \\(6 \\times 9 = 54\\) square centimeters.\n\nThus, the area of \\(\\triangle CDE\\) is 54 square centimeters.\n\n**</think>**\n\\boxed{54},

***Our* (r = 1)**

: You are a helpful assistant. Which emotion best describes the image?\nAnswer the question using one emotional category: love, joy, confusion, anger, fear, sadness, or surprise. 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{}.

<think>\nThe image depicts a peaceful scene of a cat sleeping on a cushion. The cat appears to be in a relaxed and comfortable position, with its eyes closed, suggesting a state of rest or contentment. The background includes soft, warm elements like blankets and a cozy atmosphere, which further contributes to the overall sense of tranquility. There are no signs of distress, anger, or any other negative emotions. The scene evokes a feeling of calm and serenity, which aligns most closely with the emotion of love or happiness, but the primary emotion here seems to be one of peace and contentment.\n**</think>**\n\\boxed{joy}

***Vanilla* (r = 1.0)**

<think>\nThe image depicts a cat sleeping on a cushion in a cozy room. The cat appears relaxed and comfortable, which suggests a sense of peace and contentment. There are no signs of distress or agitation that would indicate confusion, anger, fear, sadness, or surprise. The overall atmosphere of the image is calm and serene, which aligns with the emotion of love, as it evokes feelings of warmth and affection.\n**</think>**\n\\boxed{love}

***GRPO* (r = 0.1)**

<think>\nThe image depicts a serene and peaceful scene with a cat sleeping on a cushion. The setting appears cozy and calm, with soft lighting and a relaxed atmosphere. There are no signs of distress, anger, fear, sadness, or surprise. The overall mood is one of tranquility and contentment.\n**</think>**\n\\boxed{love}

***DAPO* (r = 0.1)**

<think>\nThe image depicts a cat lying on a cushion in a relaxed and comfortable position. The cat appears to be sleeping peacefully, which suggests a calm and content mood. There are no signs of distress or agitation that would indicate any of the other emotions listed. Therefore, the emotion that best describes the image is likely to be one of contentment or relaxation.\n**</think>**\n\\boxed{joy}

***Our* (r = 1)**

Figure 8: **Out-of-Domain Case Visualization**.

B.3 Visualization Analysis

We present the output cases for both in-domain (fig. 7) and out-of-domain distributions (fig. 8). For samples with high certainty, the existing GRPO leads to abnormal behavior. Both GRPO and DAPO exhibit degradation on the in-domain dataset EmoSet and the out-of-domain dataset WEBEmo in cases where the vanilla Qwen2.5-VL-7B-Instructversion could correctly answer. This suggests that the existing advantage distorts the optimization direction for samples with high trajectory certainty, ultimately leading to a performance decrease.

C Theoretical Analysis

To better understand how proposed MAPO reshapes the optimization dynamics compared with GRPO, we provide a gradient-level analysis. *Without loss of generality*, we simplify the gradient analysis by ignoring clipping and KL regularization and modeling the reward as accuracy, i.e., a Bernoulli variable. For a prompt with G rollouts and Bernoulli rewards $R_i \in \{0, 1\}$, define $A_i = R_i - \mu$ and let $p = \frac{N}{G}$, where $N = \sum_{i=1}^{G} \mathbf{1}_{\{R_i=1\}}$. Then $\mu = \frac{1}{G}\sum_i R_i = p$ and $\sigma = \sqrt{p(1-p)}$. Ignoring clipping and KL modules, the gradient of the objective is

$$\nabla_\theta \mathcal{J} = \mathbb{E}\Big[\sum_{i,t} r_{i,t}\, \hat{A}_i\, \nabla_\theta \log \pi_\theta(a_{i,t} \mid s_{i,t})\Big], \quad r_{i,t} = \frac{\pi_\theta(a_{i,t} \mid s_{i,t})}{\pi_{\text{old}}(a_{i,t} \mid s_{i,t})}. \tag{9}$$

For GRPO, the advantage is $\hat{A}_i^{\text{G}} = A_i/\sigma$.

For MAPO,

$$\hat{A}_i^{\text{M}} = (1-\lambda(p))\frac{A_i}{\sigma} + \lambda(p)\frac{A_i}{p}, \qquad \lambda(p) = 1 - 4p(1-p). \tag{10}$$

Hence, for any trajectory, we define the ratio of the gradient as:

$$\boxed{\varrho(p) \triangleq \frac{\nabla_\theta \mathcal{J}_{\text{MAPO}}}{\nabla_\theta \mathcal{J}_{\text{GRPO}}} = \frac{\hat{A}_i^{\text{M}}}{\hat{A}_i^{\text{G}}} = (1-\lambda(p)) + \lambda(p)\frac{\sigma}{p} = (1-\lambda(p)) + \lambda(p)h(p)} \tag{11}$$

where $h(p) = \sqrt{\frac{1-p}{p}}$.

Next, we analyze the property of $\varrho(p)$. Since h is smooth on $(0, 1)$ with

$$h'(p) = -\frac{1}{2p^2\, h(p)} = -\frac{1}{2\, p^{3/2}\sqrt{1-p}} < 0, \tag{12}$$

the derivative of ϱ is

$$\varrho'(p) = 4(1-2p)\big(1-h(p)\big) + \big(1-4p(1-p)\big)\, h'(p). \tag{13}$$

For $p \in (0, 1)$, we obtain that $\varrho'(p) \leq 0$, and $\varrho(\frac{1}{2}) = 1$. Thus, we have:

$$\begin{cases} \varrho(p) > 1, & p \in (0, \frac{1}{2}), \\ \varrho(p) = 1, & p = \frac{1}{2}, \\ 0 < \varrho(p) < 1, & p \in (\frac{1}{2}, 1). \end{cases} \tag{14}$$

Which implies that MAPO leads to amplified gradients than GRPO on harder samples (with $p < \frac{1}{2}$), and smaller updates on easier samples (with $p > \frac{1}{2}$). This leads to the conclusion in Sec. 3.3.