Unveiling Implicit Advantage Symmetry: Why GRPO Struggles with Exploration and Difficulty Adaptation

Zhiqi Yu*
School of Computing and Data Science
The University of Hong Kong
zhiqiyu777@gmail.com

Zhangquan Chen*
Tsinghua Shenzhen International Graduate School
Tsinghua University
czq23@mails.tsinghua.edu.cn

Mengting Liu
School of Biomedical Engineering
Sun Yat-sen University
liumt55@mail.sysu.edu.cn

Heye Zhang
School of Biomedical Engineering
Sun Yat-sen University
zhangheye@mail.sysu.edu.cn

Liangqiong Qu
School of Computing and Data Science
The University of Hong Kong
liangqqu@hku.hk

Abstract

Reinforcement Learning with Verifiable Rewards (RLVR), particularly GRPO, has become the standard for eliciting LLM reasoning. However, its efficiency in exploration and difficulty adaptation remains an open challenge. In this work, we identify an implicit advantage symmetry inherent in Group Relative Advantage Estimation (GRAE) as a structural property that provides a new perspective for understanding these bottlenecks. This symmetry induces two critical limitations: (i) at the group level, strict symmetry in weights between correct and incorrect trajectories leaves unsampled action logits unchanged, thereby hindering the exploration of novel correct solution. (ii) at the sample level, the algorithm implicitly prioritizes medium-difficulty samples, remaining agnostic to the non-stationary demands of difficulty focus. Through controlled experiments, we reveal that this symmetric property is sub-optimal, yielding two pivotal insights: (i) asymmetrically down-weighting the advantages of correct trajectories encourages essential exploration but risks instability; (ii) learning efficiency can be boosted by a curriculum-like transition—prioritizing simpler samples initially before gradually shifting to complex ones. Motivated by these findings, we propose Asymmetric GRAE (A-GRAE), which dynamically modulates exploration incentives and sample-difficulty focus. Experiments across seven benchmarks demonstrate that A-GRAE consistently improves GRPO and its variants across both LLMs and MLLMs. Code is available at https://github.com/HKU-HealthAI/A-GRAE.

1 Introduction

Reinforcement Learning with Verifiable Rewards (RLVR) [36, 1, 41, 49] has become a cornerstone for activating Foundation Models' capacity to address complex reasoning tasks through Chain-of-Thought (CoT) generation. Among various algorithms, Group Relative Policy Optimization (GRPO)

*Equal contribution.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

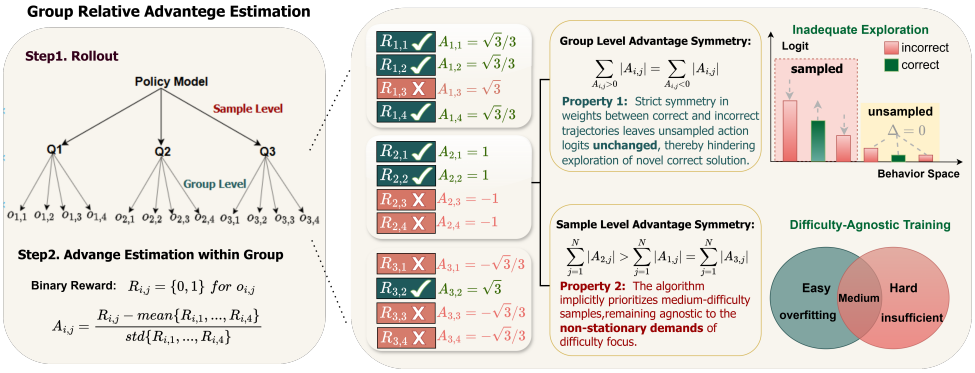

Figure 1: **The two-fold implicit advantage symmetry problem of GRAE in GRPO.** At the group level, the advantage weights for correct trajectories equal those of incorrect trajectories. This symmetry leads to the logits of low-probability correct paths unchanged under the advantage-driven update, thereby hindering the model's exploration. At the sample level, medium-difficulty samples exhibit the largest sum of absolute advantage values, leading to insufficient training on harder data.

has emerged as the standard implementation in advanced systems such as DeepSeek-R1 [16] and OpenAI-o1 [25]. The core innovation of GRPO lies in Group Relative Advantage Estimation (GRAE), which computes relative advantages within sampled groups to obviate the need for a value model.

Recent studies have identified two critical limitations of GRPO. The first is **capability boundary shrinkage**: A prominent critique posits that GRPO primarily leads to an exploration-exploitation trade-off within existing policy constraints rather than effectively expanding the decision boundary [54, 3, 45, 34, 24]. This claim is supported by empirical evidence showing that GRPO's Pass@k can even fall below that of the base model at large k. This phenomenon has motivated approaches to improves Pass@k based on negative reinforcement, entropy-aware optimization, and Pass@k-oriented objectives [62, 51, 7]. The second limitation is **inadequate focus on problem difficulty**: GRPO's reward mechanism is difficulty-agnostic, treating all tasks uniformly without accounting for their inherent complexity or the model's current capacity. To this end, recent methods[56, 61, 58, 9] have sought to improve training efficiency and accuracy by either reweighting hard or informative samples or incorporating curriculum-style strategies.

However, these two lines of work remain largely separate, and few approaches achieve substantial gains in both accuracy and Pass@k due to the fundamental conflict between sampling efficiency and diversity. This naturally raises a key question: *Can a unified perspective explain these two limitations of GRPO and, in turn, guide a principled way to overcome them at the same time?*

In this work, we show that the two deficiencies can be understood through a previously overlooked **implicit advantage symmetry inherent in GRAE**. As illustrated in Fig. 1, this symmetry appears at two complementary levels. At the group level, the advantage weights for correct and incorrect trajectories are strictly equivalent. Through a behavior-space logit analysis, we show that this symmetry implies that an unsampled trajectory receives no direct reward-specific advantage signal, thereby limiting reward-directed exploration of potentially correct but unseen solutions. At the sample level, by quantifying the absolute sum of advantages across samples, we reveal that the algorithm implicitly prioritizes medium-difficulty instances. Consequently, the optimization remains agnostic to the non-stationary demands of training dynamics, failing to adapt its focus as the model evolves.

On this theoretical ground, we perform controlled interventions that deliberately break GRAE's intrinsic advantage symmetry at both the group and sample levels to examine its causal effect on learning dynamics. Our experiments reveal that the symmetric property is sub-optimal and yield two key design principles: (i) asymmetrically down-weighting the weights of correct trajectories fosters essential exploration; (ii) learning efficiency can be boosted by a curriculum-like progression, which prioritizes simpler samples initially before escalating to complex ones. Motivated by these findings, we propose **Asymmetric GRAE (A-GRAE)**, which refines the original GRAE strategy by explicitly instantiating these two principles within the GRPO framework: it introduces asymmetric exploration to push the policy beyond its current solution set, and a curriculum-like progression learning schedule to align the optimization focus with the model's evolving capability.

To fully validate A-GRAE, we conduct comprehensive experiments across seven diverse benchmarks—spanning both natural language reasoning and vision-language reasoning tasks—using various commonly used LLMs and VLMs. Experimental results prove that A-GRAE consistently enhances the performance of GRPO and its representative variants DAPO [53] and Dr.GRPO [31] across all settings, with significant improvements in key metrics (e.g., accuracy, Pass@k).

In summary, our contributions are as follows:

- We identify **advantage symmetry** as a two-level structural property of GRAE, providing a unified perspective that connects previously separate methods for exploration and difficulty-aware learning in GRPO-based RLVR.
- Based on this unified perspective, we systematically break group-level and sample-level symmetry through controlled experiments and uncover two distinct findings beyond existing fixed strategies: excessive group-level exploration can cause training instability, while the preferred sample difficulty dynamically shifts from easier to harder queries during training.
- Guided by these findings, we propose **A-GRAE**, which jointly and dynamically adjusts group-level exploration and sample-level difficulty focus. Extensive experiments across seven benchmarks, multiple model families, and several GRPO variants demonstrate consistent improvements in both accuracy and Pass@k.

2 Preliminary

Notation. In this work, we frame a foundation model (LLM or MLLM), parameterized by θ, as a policy model: π_θ and $\pi_{\theta_{\text{old}}}$ denote the current and prior policies, respectively. For batch processing, we sample a question q from the dataset $\mathcal{Q}$. For each question q, the policy generates a set of G candidate responses $\{o_i\}_{i=1}^{G}$. Each question-response pair (q, o_i) is then assigned a scalar reward r_i via a rule-based verifier, where $i \in \{1, 2, \ldots, G\}$ indexes responses per question. By default, we employ a binary accuracy reward: $r_i = 1$ if response o_i is correct, and $r_i = 0$ otherwise. For each response o_i, we further compute a group-relative advantage value A_i for policy optimization.

Group Relative Policy Optimization (GRPO). GRPO simplifies the training pipeline of PPO by removing the critic model—typically matching the size of the policy model—and instead estimates baselines via group-level reward scores. Specifically, for each query q, GRPO samples a set of responses $\{o_1, o_2, \ldots, o_G\}$ from the prior policy $\pi_{\theta_{\text{old}}}$ and optimizes the target policy π_θ by maximizing the following objective function:

$$J_{\text{GRPO}}(\pi_\theta) = \mathbb{E}_{q\sim\mathcal{Q},\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q)} \frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\Big[\min\big(\rho_{i,t}A_{i,t}, \text{clip}\left(\rho_{i,t}, 1\pm\epsilon\right)A_{i,t}\big) - \beta\mathbb{D}_{\text{KL}}\Big], \tag{1}$$

where $\rho_{i,t} = \frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q,o_{i,<t})}$, ϵ and β are hyperparameters, the KL term is defined as

$$\mathbb{D}_{KL} = \frac{\pi_{ref}(o_{i,t}|q,o_{i,<t})}{\pi_\theta(o_{i,t}|q,o_{i,<t})} - \log\frac{\pi_{ref}(o_{i,t}|q,o_{i,<t})}{\pi_\theta(o_{i,t}|q,o_{i,<t})} - 1, \tag{2}$$

and the advantage A_i is computed using a group of rewards $\{r_1, r_2, ..., r_G\}$ with GRAE:

$$A_i = \frac{r_i - \text{mean}(\{r_1, r_2, ..., r_G\})}{\text{std}(\{r_1, r_2, ..., r_G\})}. \tag{3}$$

3 Implicit Advantage Symmetry in GRPO

First, we demonstrate that GRPO's policy gradient optimization can be formulated as **a reweighting variant of Supervised Fine-Tuning (SFT)**. Specifically, the gradient update can be formulated as

follows (proof in Section B.1):

$$\nabla_\theta \mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim\mathcal{Q},\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q)} \underbrace{\frac{1}{G}\sum_{i=1}^{G}\left[\underbrace{\frac{1}{|\mathbf{o}_i|}\sum_{t=1}^{|\mathbf{o}_i|}\underbrace{\rho_{i,t}A_{i,t}}_{\text{weight}}\underbrace{\nabla_\theta \log \pi_\theta(o_{i,t} \mid q, \mathbf{o}_{i,<t})}_{\text{SFT}}}_{\text{Group Level}}\right]}_{\text{Sample Level}}. \tag{4}$$

In Eq. (4), $\rho_{i,t}$ is relatively stable with the clip operation, so the dominant part of reweighting is $A_{i,t}$. In general, the advantage of GRPO is shared across the entire sequence, meaning that all tokens within a single response correspond to the same advantage value, i.e.,$A_{i,1} = A_{i,2} = \cdots = A_{i,t} = A_i$. We then examine this reweighting mechanism from the group level and the sample level.

Advantage Symmetry in Group Level: In RLVR, a trajectory is considered correct provided that the final answer extracted from the response aligns with the ground truth. Given a query, we partition the sampled within-group responses into correct (positive) and incorrect (negative) trajectories. Crucially, we observe that the aggregate weights of policy updates attributed to correct trajectories are equivalent to those attributed to incorrect ones (proof in Section B.2). This property can be expressed as:

$$\sum_{i\in G_{pos}} |A_i| = \sum_{i\in G_{neg}} |A_i| \tag{5}$$

where G_{pos} and G_{neg} represent the sets of positive and negative trajectories, satisfying the conditions $G_{pos} \cap G_{neg} = \emptyset$ and $G_{pos} \cup G_{neg} = G$ (the sampled response set). This property is subsequently evaluated within the behavior space.

Theorem 1 (Behavior-Space Logit Update) *Assume that the set of all possible behaviors is $B = \{b_i\}_{i=1}^{N}$, consisting of sampled set G and unsampled set U. Given the intragroup advantage sum $C = \sum_{b_i\in G} A_i$, the behavior-level logit update of b_i under gradient ascent can be expressed as:*

$$\Delta h_{b_i} = \eta \left[\mathbb{I}(b_i \in G)A_{b_i} - C\pi_{b_i}\right]. \tag{6}$$

where η denotes the learning rate and π_{b_i} is the model's current sampling probability in behavior space for trajectory b_i (proof in Section B.3).

In standard GRAE, advantage normalization enforces the zero-sum property $C = 0$. Therefore, sampled correct trajectories are directly reinforced, while sampled incorrect trajectories are directly suppressed. In contrast, an unsampled trajectory receives no direct advantage-driven update in the behavior-space abstraction. Moreover, for a parameterized nonlinear policy, shared parameters still indirectly change the probability of an unsampled trajectory as our parameter-space analysis in Section B.3 shows that such changes arise only through gradient alignment with sampled trajectories and contain no reward signal specific to the unsampled trajectory.

Advantage Symmetry in Sample Level: At the sample level, the overall learning contribution can be captured by the sum of absolute advantages, which represents the total weights of one query. To evaluate the relative contributions across queries of varying complexities, we introduce the sample success rate, denoted as p for a group of size $|G|$, as a proxy for task difficulty; specifically, a higher p indicates a lower level of difficulty. By leveraging this metric, we can formally quantify the relationship between sample difficulty and the corresponding update magnitude.

Theorem 2 (Update Magnitude with respect to Sample Difficulty). *Consider a group G of trajectories for one query (sample) with binary rewards $\{r_i\}_{i=1}^{G}$, the sum of absolute advantages over the group under Group Relative Advantage Estimation (GRAE) can be derived as:*

$$\sum_{i\in G} |A_i| = 2|G|\sqrt{p(1-p)}. \tag{7}$$

where $p = \sum_{i\in G} |r_i|/|G|$ denotes the empirical success probability of the corresponding sample within group G, and the proof can be found in Section B.4.

Theorem 2 reveals an intrinsic difficulty bias in GRAE: samples of intermediate difficulty ($p = 0.5$) dominate policy updates. As illustrated in Fig. 4, the symmetry of $\sqrt{p(1-p)}$ assigns identical

weights to easy ($p = 0.75$) and hard ($p = 0.25$) samples. This static approach, however, ignores the dynamic distributional shift during training, where simple instances increasingly dominate the sample pool. Such symmetry causes the model to overfit on trivial data at the expense of challenging scenarios, failing to satisfy the non-stationary demands of the evolving learning process.

4 Deconstructing the Implicit Advantage Symmetry of GRAE

4.1 Experimental Setup

Table 1: **Summary of Controlled Experiments.** We design two experiments to investigate the impact of symmetry at the group and sample levels. $\beta = 10$ and $\gamma = 0.5$ are scaling constants.

(a) **Exp I: Breaking Group-Level Symmetry.**

Variant	**Formulation** (A^*)	$\sum_i^G A_i^*$
GRPO (Control Group)	$A_{\text{pos}}^* = A_{\text{pos}}$	$= 0$
Positive-Dominant	$A_{\text{pos}}^* = \beta \cdot A_{\text{pos}}$	> 0
Negative-Dominant	$A_{\text{pos}}^* = A_{\text{pos}}/\beta$	< 0

(b) **Exp II: Breaking Sample-Level Symmetry.**

Variant	**Formulation** (A^*)	$\sum_i^G \lvert A_i^* \rvert$
GRPO (Control Group)	$A_i^* = A_i$	$2\lvert G \rvert \sqrt{p(1-p)}$
Hard-Focused	$A_i^* = \gamma \cdot A_i/\sqrt{p}$	$\lvert G \rvert \sqrt{1-p}$
Easy-Focused	$A_i^* = \gamma \cdot A_i/\sqrt{1-p}$	$\lvert G \rvert \sqrt{p}$

To systematically investigate the impact of advantage symmetry on reasoning performance, we design two sets of ablation experiments as shown in Table 1:

Exp I: Breaking Group-Level Symmetry. This experiment isolates the effect of violating the group-level symmetry by asymmetrically scaling the advantages of correct trajectories. To achieve this, we introduce a scaling coefficient $\beta = 10$ to disrupt the zero-sum equilibrium ($\sum A_i = 0$). Denoting the original positive advantage as A_{pos}, we design two variants:

1. *Positive-Dominant Group:* We scale up positive advantages ($A_{\text{pos}}^* = \beta \cdot A_{\text{pos}}$).
2. *Negative-Dominant Group:* We scale down positive advantages ($A_{\text{pos}}^* = A_{\text{pos}}/\beta$).

The original GRPO serves as the control group, representing the symmetric equilibrium.

Exp II: Breaking Sample-Level Symmetry. This experiment investigates difficulty focus during training by modifying advantage magnitudes using the sampling success rate p. Let A_i be the original advantage of GRPO. We define two curriculum variants:

1. *Hard-Focused Group:* We shift the learning focus toward harder queries by rescaling the advantage with the factor $1/\sqrt{p}$ ($A_i^* = \gamma \cdot A_i/\sqrt{p}$).
2. *Easy-Focused Group:* Conversely, we shift the focus to simple queries by scaling with the factor $1/\sqrt{1-p}$ ($A_i^* = \gamma \cdot A_i/\sqrt{1-p}$).

Here, $\gamma = 0.5$ is a normalization constant to ensure that the theoretical maximum value remains consistent with the control group (standard GRPO). Note that no extra rescaling is conducted when the success rate is 0 or 1, as the GRPO advantage is zero in these cases.

Models and Training Setup: We conducte experiments on math-specialized `Qwen2.5-Math-7B` [50] and the general-purpose `Llama-3.2-3B-Instruct` [12]. All models are trained on the MATH dataset [20] using `verl` framework [43]. Key configurations include a batch size of 1,024, $G = 8$ rollouts per query, and a learning rate of 1×10^{-6}. We evaluate model performance across MATH, AMC23, and the high-difficulty AIME 2025 benchmarks using an unbiased Pass@k estimator, consistent with established literature [21, 5, 7, 62]. Other training details are available in Section C.2.

4.2 Experiments on Breaking Group Level Symmetry

GRPO enhances efficiency without expanding reasoning boundaries. As shown in Fig. 2, while GRPO significantly improves Pass@1 across all datasets, these gains erode at higher sampling budgets. Notably, on the AMC23 and MATH datasets, GRPO's performance at Pass@256 falls below that of the base model. This suggests that while GRPO improves the sampling probability of correct paths, it fails to discover novel solutions that lie outside the base model's original sampling support. In other words, GRPO does not fundamentally expand the intrinsic reasoning boundaries of the base model. This observation aligns with findings in recent literature [54, 51].

Over-emphasizing correct paths triggers entropy collapse. In contrast to GRPO, amplifying the weight of positive trajectories fails to yield performance improvements. In particular, on the

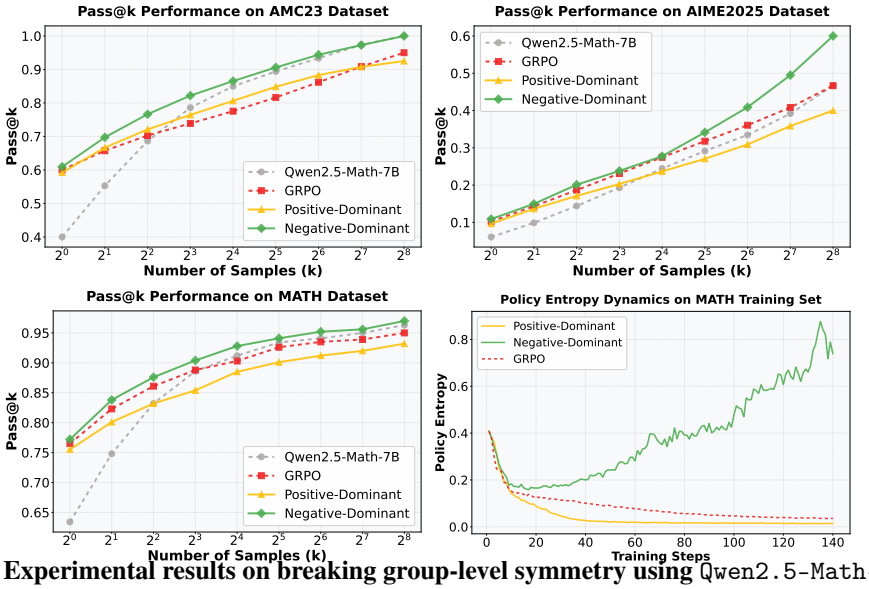

Figure 2: **Experimental results on breaking group-level symmetry using `Qwen2.5-Math-7B`.** We amplify (Positive-Dominant) or suppress (Negative-Dominant) the advantages of correct trajectories to compare their performance with those of GRPO and the base model.The performance is evaluated using Pass@k. The last figure illustrates the entropy dynamics on the training set.

AIME2025 and MATH datasets, the Positive-Dominant group significantly underperforms other methods at larger sampling budgets ($k > 16$). To investigate the underlying cause, we find that the policy entropy measured on the training set for the Positive-Dominant group exhibits the most precipitous decline, as illustrated in Fig. 2. This phenomenon indicates that overemphasis on correct trajectories reduces sampling diversity by excessively sharpening the output distribution, leading to entropy collapse and thus impairing the model's reasoning capability.

Down-weighting the correct path improves performance but risks instability. Notably, down-weighting positive trajectories consistently outperforms vanilla GRPO in Pass@k metrics across all datasets. Even at a high sampling budget ($k = 256$), this Negative-Dominant approach maintains parity with the base model, suggesting it effectively mitigates the capability boundary shrinkage inherent in GRPO. As shown in Fig. 2, this improvement correlates with a sustained increase in policy entropy, indicating enhanced exploration. Importantly, this gain is not merely a byproduct of generic entropy inflation. As demonstrated in our entropy-matched control experiments (Section D.5), naively promoting entropy degrades performance and induces repetitive outputs. Rather, what matters is how exploration is promoted at the trajectory level: inherently breaking advantage symmetry prevents indiscriminate broad support diffusion and fosters meaningful diversity. However, in practice, the Negative-Dominant strategy can trigger training instability (demonstrated in Section D.3) when overconfident incorrect trajectories displace correct ones (discussed in Section B.5). Consequently, a dynamic adjustment mechanism is essential to balance exploration diversity with training stability as training progresses.

> ***Takeaway 1:*** *As discussed in Section 4.2, while down-weighting positive advantage incentivizes the model to explore unsampled correct trajectories to improve the reasoning ability, it also introduces the risk of training collapse in the later stages of reinforcement learning.*

4.3 Experiments on Breaking Sample Level Symmetry

Static difficulty reweighting is not universally optimal and depends on test difficulty. As shown in Fig. 3, none of the methods yielded a uniform advantage across all datasets. More precisely, their relative efficacy varies by benchmark: hard-focused methods achieve peak results on the challenging AIME2025, whereas easy-focused approaches marginally outperform others on simpler datasets such as AMC23 and MATH at Pass@1. These findings indicate the optimal difficulty weighting is

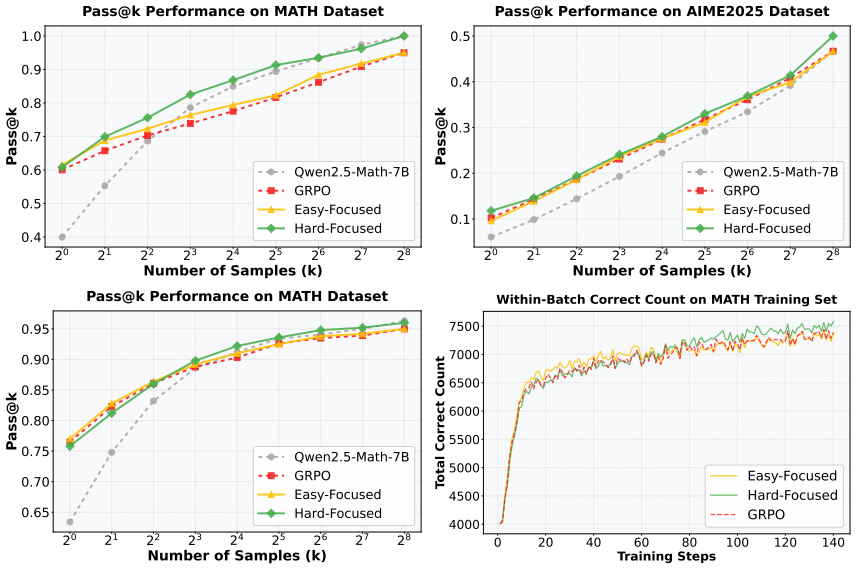

Figure 3: **Experimental results on breaking sample-level symmetry using `Qwen2.5-Math-7B`.** We rescale the advantages to shift the learning focus toward harder queries (Hard-Focused) or easier queries (Easy-Focused) to compare their performance with those of GRPO and the base model.The performance is evaluated using Pass@k. The last figure illustrates the within-batch count of correct sampled responses on the training set.

evaluation-dependent, challenging the prevailing trend of focusing solely on hard samples throughout training [56, 39, 15].

Prioritizing simple samples in early stages and hard samples later yields better learning efficiency gains. As direct evaluation fails to identify a universally optimal strategy, we further investigate the training dynamics. As shown in Fig. 3, the Easy-Focused strategy accelerates most rapidly in early phases, which suggests that prioritizing simpler tasks promotes the learning of basic formatting rules and core reasoning patterns. Conversely, the Hard-Focused strategy excels in later stages. This shift suggests that once model capability reaches relative saturation, a transition toward difficult samples becomes essential to further elevate the performance ceiling. Unlike prior curriculum RL methods that manually impose fixed "easy-to-hard" schedules based on predefined rankings [6, 13] or focus solely on medium-difficulty or difficult samples [55, 2], our findings reveal a non-stationary learning pattern that yields better learning efficiency gains.

> ***Takeaway 2:*** *Our findings in Section 4.3 suggest that static difficulty reweighting is insufficient, as optimal sample utility is phase-dependent. While easy samples facilitate more effective learning in the early stage, a strategic transition toward hard samples becomes useful as performance begins to plateau.*

4.4 A Unified Perspective on Existing RL Methods

The proposed **advantage symmetry** provides a unified perspective for connecting existing GRPO-based methods that were originally developed from different motivations. From this perspective, negative-reinforcement and Pass@k-oriented approaches [62, 51, 35, 29, 47] can be interpreted as implicitly breaking **group-level symmetry** to promote exploration, whereas methods that emphasize high-entropy tokens or difficult samples [8, 17, 60, 11, 38, 56, 42, 55, 2] mainly break **sample-level symmetry** by reallocating learning signals toward uncertain or hard regions.

Consequently, this unified formulation enables controlled interventions along the two symmetry axes and reveals findings beyond existing fixed asymmetric strategies. At the group level, stronger exploration induced by suppressing positive advantages is beneficial but can become unstable when maintained excessively; at the sample level, a fixed preference for hard samples is not always optimal, and the preferred difficulty shifts from easier queries in early training to harder ones in later stages.

Table 2: Pass@k results on MATH, AIME 2025 and AMC23 with `Qwen2.5-Math-7B`. **Bold** and underlined numbers denote the best and second-best results for each k. Specifically, We conduct 16 runs on AIME2025 and AMC23, and 4 runs on the remaining benchmarks.

Method	**Pass@k**									
k	1	2	4	8	16	32	64	128	256	**Avg.**
MATH										
Base Model	63.4 ±0.4	74.8 ±0.5	83.2 ±0.6	88.6 ±0.7	91.2 ±0.7	93.4 ±0.6	94.1 ±0.5	95.0 ±0.4	96.3 ±0.3	86.7
SEC	77.6 ±0.5	82.5 ±0.6	86.0 ±0.6	88.9 ±0.4	90.2 ±0.5	92.2 ±0.6	92.6 ±0.5	93.3 ±0.3	94.7 ±0.3	88.7
GRPO-LEAD	77.8 ±0.3	83.0 ±0.4	86.5 ±0.5	89.2 ±0.5	90.5 ±0.5	92.3 ±0.5	92.8 ±0.4	93.6 ±0.4	95.0 ±0.3	89.0
P@kT	76.2 ±0.6	82.2 ±0.8	86.9 ±0.6	89.6 ±0.4	91.4 ±0.3	93.3 ±0.4	94.0 ±0.5	96.2 ±0.4	96.2 ±0.4	89.3
W-REINFORCE	76.6 ±0.5	82.8 ±0.6	87.1 ±0.7	90.2 ±0.8	92.4 ±0.8	94.1 ±0.7	95.3 ±0.6	96.1 ±0.5	96.7 ±0.4	90.1
GRPO	76.5 ±0.3	82.3 ±0.4	86.1 ±0.5	88.8 ±0.5	90.3 ±0.5	92.6 ±0.5	93.5 ±0.4	93.9 ±0.4	95.0 ±0.3	88.8
GRPO + A-GRAE	78.3 ±0.2	85.0 ±0.3	89.2 ±0.4	**91.0** ±0.4	92.5 ±0.4	94.6 ±0.4	95.0 ±0.3	95.5 ±0.3	96.5 ±0.2	90.8
DAPO	75.0 ±0.4	79.8 ±0.5	85.0 ±0.6	88.4 ±0.6	89.8 ±0.6	92.0 ±0.6	92.8 ±0.5	93.4 ±0.4	94.3 ±0.4	87.8
DAPO + A-GRAE	76.9 ±0.3	82.6 ±0.4	86.5 ±0.5	89.2 ±0.5	90.6 ±0.5	92.8 ±0.5	93.6 ±0.4	94.2 ±0.4	95.3 ±0.3	89.1
Dr.GRPO	77.2 ±0.4	82.5 ±0.5	87.4 ±0.6	89.6 ±0.6	91.0 ±0.6	92.8 ±0.5	93.6 ±0.5	94.3 ±0.4	95.0 ±0.4	89.3
Dr.GRPO + A-GRAE	**78.6** ±0.3	**86.2** ±0.3	**89.8** ±0.4	90.6 ±0.4	**92.8** ±0.4	**95.0** ±0.4	**95.4** ±0.3	**96.0** ±0.3	**96.9** ±0.2	**91.3**
AIME 2025										
Base Model	6.1 ±0.7	9.9 ±0.9	14.4 ±1.1	19.3 ±1.3	24.4 ±1.5	29.1 ±1.6	33.4 ±1.6	39.2 ±1.5	46.7 ±1.3	24.7
SEC	11.3 ±0.8	15.2 ±1.0	19.5 ±0.9	23.8 ±1.2	28.3 ±1.5	32.1 ±1.8	36.4 ±1.6	41.0 ±1.3	49.0 ±1.2	28.5
GRPO-LEAD	11.0 ±0.5	14.8 ±0.7	19.2 ±0.9	23.4 ±1.1	27.8 ±1.3	32.0 ±1.4	36.5 ±1.5	41.4 ±1.4	47.3 ±1.3	28.2
P@kT	10.8 ±0.5	15.6 ±0.6	20.6 ±1.2	25.0 ±1.4	29.8 ±1.3	34.5 ±1.5	40.3 ±1.4	47.6 ±1.6	56.4 ±1.5	31.2
W-REINFORCE	10.6 ±0.7	15.3 ±0.9	20.0 ±1.1	24.7 ±1.3	29.7 ±1.5	34.6 ±1.6	40.5 ±1.7	47.8 ±1.6	56.7 ±1.4	31.1
GRPO	10.3 ±0.5	14.3 ±0.7	18.7 ±0.9	23.1 ±1.1	27.5 ±1.3	31.8 ±1.4	36.1 ±1.5	40.8 ±1.4	46.7 ±1.3	27.7
GRPO + A-GRAE	11.3 ±0.4	15.6 ±0.6	19.8 ±0.8	24.7 ±1.0	28.6 ±1.2	34.1 ±1.3	39.2 ±1.3	47.8 ±1.2	56.7 ±1.1	30.9
DAPO	12.0 ±0.6	16.1 ±0.8	21.3 ±1.0	25.2 ±1.2	29.4 ±1.4	33.2 ±1.5	38.5 ±1.5	45.4 ±1.4	53.3 ±1.2	30.5
DAPO + A-GRAE	**13.3** ±0.4	**18.4** ±0.6	**23.0** ±0.8	**26.3** ±1.0	**30.0** ±1.2	**35.1** ±1.3	**41.1** ±1.3	**48.7** ±1.2	**60.0** ±1.0	**32.9**
Dr.GRPO	11.0 ±0.6	14.8 ±0.8	19.3 ±1.0	24.3 ±1.2	28.8 ±1.4	33.0 ±1.5	37.1 ±1.5	41.2 ±1.4	46.7 ±1.3	28.5
Dr.GRPO + A-GRAE	11.8 ±0.4	16.2 ±0.6	19.8 ±0.8	25.0 ±1.0	29.3 ±1.2	34.8 ±1.3	37.9 ±1.3	48.0 ±1.2	56.7 ±1.1	31.1
AMC23										
Base Model	40.6 ±0.5	55.3 ±0.6	68.6 ±0.8	78.6 ±0.9	85.0 ±0.9	89.4 ±0.8	93.4 ±0.7	**97.3** ±0.5	**100.0** ±0.0	78.7
SEC	61.8 ±0.5	67.5 ±0.8	72.8 ±0.7	77.5 ±1.0	81.0 ±0.8	84.6 ±0.6	87.8 ±0.6	90.2 ±0.6	92.1 ±0.5	79.5
GRPO-LEAD	62.3 ±0.4	68.0 ±0.5	73.3 ±0.6	77.8 ±0.7	81.5 ±0.8	85.0 ±0.7	88.2 ±0.6	90.3 ±0.5	92.3 ±0.4	79.9
P@kT	61.3 ±0.5	68.8 ±0.8	76.0 ±0.8	82.3 ±1.0	87.3 ±0.8	91.6 ±0.7	94.8 ±0.6	96.7 ±0.5	97.2 ±0.6	84.0
W-REINFORCE	62.0 ±0.6	70.0 ±0.7	77.0 ±0.8	83.1 ±0.9	87.8 ±0.9	91.8 ±0.8	95.0 ±0.7	97.0 ±0.5	97.5 ±0.4	84.6
GRPO	60.2 ±0.4	66.7 ±0.5	72.1 ±0.6	76.4 ±0.7	80.6 ±0.7	84.8 ±0.7	88.3 ±0.6	90.8 ±0.5	92.5 ±0.4	79.2
GRPO + A-GRAE	62.6 ±0.3	70.0 ±0.4	77.5 ±0.5	83.7 ±0.6	88.2 ±0.7	92.0 ±0.7	95.1 ±0.6	96.8 ±0.4	97.5 ±0.3	84.8
DAPO	62.0 ±0.4	70.3 ±0.5	77.2 ±0.6	83.1 ±0.7	87.8 ±0.8	91.4 ±0.8	94.0 ±0.7	96.1 ±0.5	97.5 ±0.4	84.4
DAPO + A-GRAE	**63.3** ±0.3	**72.5** ±0.4	**80.5** ±0.5	**86.7** ±0.6	**90.2** ±0.7	**92.9** ±0.7	95.0 ±0.6	97.0 ±0.4	**100.0** ±0.0	**86.5**
Dr.GRPO	60.7 ±0.5	69.8 ±0.6	75.6 ±0.7	82.6 ±0.8	87.8 ±0.8	90.9 ±0.8	93.2 ±0.7	94.6 ±0.6	95.0 ±0.5	83.4
Dr.GRPO + A-GRAE	62.8 ±0.3	71.6 ±0.4	78.2 ±0.5	84.0 ±0.6	89.6 ±0.7	92.3 ±0.7	**95.2** ±0.6	95.9 ±0.5	**100.0** ±0.0	85.5

These observations suggest that the two forms of symmetry should be **jointly and dynamically broken according to the training state**, which directly motivates the design of A-GRAE[2].

5 Asymmetric GRAE

5.1 Method

Our prior analysis reveals that the advantage symmetry inherent in GRPO undermines model exploration and difficulty adaptation. To address these limitations, we propose the Asymmetric Group Relative Advantage Estimation (A-GRAE) framework to dynamically modulate exploration incentives and sample-difficulty focus. To implement this, a metric is required to quantify training state; accordingly, we introduce the batch-wise mean reward as a proxy indicator:

$$\omega_s = \frac{\sum_{i=1}^{B} r_i}{B}, \tag{8}$$

where B denotes the total number of trajectories in the batch, ω_s denotes the mean reward of the current step, where a higher value implies stronger model proficiency. Then we introduce a dynamic attention shift at the sample level, transitioning from easy to hard samples as training progresses:

$$A_i = \frac{\omega_s}{2} \cdot \frac{r_i - \text{mean}(\{r_1, r_2, ..., r_G\})}{\text{std}(\{r_1, r_2, ..., r_G\}) \cdot \sqrt{p}} + \frac{1-\omega_s}{2} \cdot \frac{r_i - \text{mean}(\{r_1, r_2, ..., r_G\})}{\text{std}(\{r_1, r_2, ..., r_G\}) \cdot \sqrt{1-p}}, \tag{9}$$

where p denotes the sampling success rate for a given query. Note that rescaling is omitted when $p \in \{0, 1\}$, as the standard advantage equals 0. As the model evolves with increasing sampling success rate, the weight assigned to the hard-focused component (first term) progressively increases, while the corresponding weight for the easy-focused component (second term) diminishes. This mechanism facilitates adaptive trade-offs, dynamically shifting the training focus toward harder questions as the model's proficiency improves.

[2]More related work is discussed in detail in Section A.

At the group level, we propose an attenuation suppression strategy for correct-response advantages, which encourages exploration in the early stage while preserving stability in the later phase:

$$A_i^* = \begin{cases} A_i * \min(1, \frac{\omega_s}{\alpha}), & \text{if } A_i > 0 \\ A_i, & \text{if } A_i \leq 0 \end{cases} \tag{10}$$

Here, $\alpha > 0$ is a scaling parameter. Once the refined advantages are computed, they can be seamlessly incorporated into the GRPO function, as defined in Eq. (1), or other GRPO variants for optimization.

5.2 Experimental Setup

Table 3: Results on multi-modal benchmarks with `Qwen2.5-VL-3B-Instruct`. We conduct 4 runs on each benchmark. **Bold** and underlined numbers denote the best and second-best results.

	ID Domain	**OOD Domain**	
Method	Geo3K	MathVision	Mathverse
Task A: General Mathematical Reasoning			
Base Model	27.8 ±1.0	20.8 ±0.8	31.6 ±0.7
GRPO	43.5 ±0.8	23.4 ±0.7	35.2 ±0.9
GRPO + A-GRAE	45.7 ±0.6	24.0 ±0.4	36.8 ±0.5
DAPO	44.7 ±0.9	23.8 ±0.6	35.9 ±0.8
DAPO + A-GRAE	45.9 ±0.4	24.3 ±0.5	37.5 ±0.8
Dr.GRPO	44.9 ±0.7	24.2 ±0.8	36.5 ±0.7
Dr.GRPO + A-GRAE	**46.8** ±0.5	**25.6** ±0.6	**38.4** ±0.6
Task B: Medical Imaging Reasoning			
	MRI300	CT300	Xray300
Base Model	35.6 ±1.6	42.5 ±2.1	42.0 ±1.4
GRPO	87.2 ±1.2	71.7 ±0.9	63.2 ±1.1
GRPO + A-GRAE	88.2 ±0.8	73.1 ±0.7	71.3 ±1.0
DAPO	84.3 ±1.1	71.6 ±1.0	63.1 ±0.9
DAPO + A-GRAE	87.0 ±1.0	72.3 ±0.8	71.6 ±0.7
Dr.GRPO	87.8 ±0.9	72.4 ±0.8	69.5 ±0.7
Dr.GRPO + A-GRAE	**89.0** ±0.6	**73.6** ±0.7	**72.0** ±0.5

We validate our method across seven benchmarks, including text-only mathematical tasks (AIME2025, AMC23, MATH [20]) with `Qwen2.5-Math-7B` and `DeepSeek-R1-7B` (Section D.2) and multimodal tasks in mathematics (Geo3K [33], MathVision [46], MathVerse [57]) and medicine (HuatuoGPT-Vision [4]) with `Qwen2.5-VL-3B-Instruct`. Our experiments evaluate GRPO variants (GRPO, DAPO, Dr.GRPO) alongside representative baselines. These include pass@k optimization methods targeting exploration efficiency (W-REINFORCE [62], p@kT [7]) and curriculum learning approaches for difficulty adaptation (GRPO-LEAD [56], SEC [6]). We report Pass@k on the text datasets and Pass@1 on the multimodal datasets since most of them are multiple-choice questions. For details please refer to Section C.2.

5.3 Main Results

Consistent improvements in Pass@1 and Pass@k. As illustrated in Table 2, our proposed method yields substantial performance gains when integrated with various GRPO variants, which proves the effectiveness of our proposed method (the statistical significance tests can be found in Section D.6). Notably, existing baselines often exhibit a trade-off between accuracy and diversity. For instance, curriculum learning approaches SEC and GRPO-LEAD primarily improves Pass@1 but shows limited gains in Pass@k categories. Conversely, pass@k optimization methods W-REINFORCE and P@kT emphasize Pass@k performance while achieving relatively modest improvements in greedy accuracy. In contrast, our A-GRAE realizes consistent gains across both metrics, which proves the superiority of our method.

Universal applicability to multi-modal domains. To evaluate the versatility of A-GRAE, we extended our empirical analysis to the multimodal domain, with the results summarized in Table 3. The results reveal that A-GRAE not only delivers significant improvements in the in-distribution (ID) domain but also yields substantial gains in out-of-distribution (OOD) scenarios. These findings provide strong evidence that our approach effectively enhances sampling efficiency while simultaneously preserving the model's generalization capabilities. In summary, the efficacy of A-GRAE across diverse domains and tasks underscores its universal applicability as a robust framework.

5.4 Further Analysis

Ablation studies. To validate the contribution of each component in A-GRAE, we perform a series of ablation studies, and the results are demonstrated in Table 5. Firstly, sample-level asymmetry primarily bolsters Pass@1 performance, consistently attaining second-best results across all three benchmarks. Conversely, the group-level asymmetric mechanism exerts a more pronounced impact on enhancing

Pass@k metrics, suggesting that this module effectively fosters generative diversity. Ultimately, the full framework yields the most significant gains across all evaluation metrics, demonstrating that the constituent modules of our methodology are mutually complementary.

Training dynamics. Figure 8 illustrates the training dynamics, specifically entropy and greedy decoding accuracy on the MATH dataset. While GRPO suffers from a continuous entropy decline on the training set, A-GRAE maintains a sustained plateau after an initial drop, successfully averting entropy collapse. On the test set, A-GRAE follows a "rise-then-fall" entropy trajectory, characteristic of a balanced exploration-exploitation paradigm. This dynamic enables A-GRAE to significantly outpace GRPO in greedy decoding accuracy during the latter training stages, confirming its superior capability in fostering sustained learning.

6 Conclusion

In this paper, we rethink the mechanics of Group Relative Advantage Estimation (GRAE) within the GRPO framework. Our theoretical and empirical analyses reveal a previously overlooked “advantage symmetry" in standard GRPO, and we demonstrate that this symmetry restricts exploration and fails to adapt its difficulty focus during the learning process. To address these limitations, we propose A-GRAE, a novel mechanism that dynamically modulates exploration incentives and prioritizes samples based on their evolving utility. Extensive evaluations across seven benchmarks demonstrate the effectiveness and superiority of A-GRAE.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (62306253), the Early Career Fund (27207025), and the Guangdong Natural Science FundGeneral Program (2024A1515010233).

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Sanghwan Bae, Jiwoo Hong, Min Young Lee, Hanbyul Kim, JeongYeon Nam, and Donghyun Kwak. Online difficulty filtering for reasoning oriented reinforcement learning. *arXiv preprint arXiv:2504.03380*, 2025.

[3] Udbhav Bamba, Minghao Fang, Yifan Yu, Haizhong Zheng, and Fan Lai. Xrpo: Pushing the limits of grpo with targeted exploration and exploitation. *arXiv preprint arXiv:2510.06672*, 2025.

[4] Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, et al. Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. *arXiv preprint arXiv:2406.19280*, 2024.

[5] Mark Chen. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

[6] Xiaoyin Chen, Jiarui Lu, Minsu Kim, Dinghuai Zhang, Jian Tang, Alexandre Piché, Nicolas Gontier, Yoshua Bengio, and Ehsan Kamalloo. Self-evolving curriculum for llm reasoning. *arXiv preprint arXiv:2505.14970*, 2025.

[7] Zhipeng Chen, Xiaobo Qin, Youbin Wu, Yue Ling, Qinghao Ye, Wayne Xin Zhao, and Guang Shi. Pass@k training for adaptively balancing exploration and exploitation in large reasoning models. *arXiv preprint arXiv:2508.10751*, 2025.

[8] Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Wayne Xin Zhao, Zhenliang Zhang, and Furu Wei. Reasoning with exploration: An entropy perspective. *arXiv preprint arXiv:2506.14758*, 2025.

[9] Yanqi Dai, Yuxiang Ji, Xiao Zhang, Yong Wang, Xiangxiang Chu, and Zhiwu Lu. Harder is better: Boosting mathematical reasoning via difficulty-aware grpo and multi-aspect question reformulation. *arXiv preprint arXiv:2601.20614*, 2026.

[10] Xingyu Dang, Christina Baek, J Zico Kolter, and Aditi Raghunathan. Assessing diversity collapse in reasoning. In *Scaling Self-Improving Foundation Models without Human Supervision*, 2025.

[11] Jia Deng, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, and Ji-Rong Wen. Decomposing the entropy-performance exchange: The missing keys to unlocking effective reinforcement learning. *arXiv preprint arXiv:2508.02260*, 2025.

[12] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407, 2024.

[13] Thomas Foster, Anya Sims, Johannes Forkel, and Jakob Nicolaus Foerster. Lilo: Learning to reason at the frontier of learnability. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[14] Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D Goodman. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars. *arXiv preprint arXiv:2503.01307*, 2025.

[15] Wei Guan, Jun Lan, Jian Cao, Hao Tan, Huijia Zhu, and Weiqiang Wang. Emit: Enhancing mllms for industrial anomaly detection via difficulty-aware grpo. *arXiv preprint arXiv:2507.21619*, 2025.

[16] 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.

[17] Zhezheng Hao, Hong Wang, Haoyang Liu, Jian Luo, Jiarui Yu, Hande Dong, Qiang Lin, Can Wang, and Jiawei Chen. Rethinking entropy interventions in rlvr: An entropy change perspective. *arXiv preprint arXiv:2510.10150*, 2025.

[18] Andre Wang He, Daniel Fried, and Sean Welleck. Rewarding the unlikely: Lifting grpo beyond distribution sharpening. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25559–25571, 2025.

[19] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

[20] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

[21] Andreas Hochlehnert, Hardik Bhatnagar, Vishaal Udandarao, Samuel Albanie, Ameya Prabhu, and Matthias Bethge. A sober look at progress in language model reasoning: Pitfalls and paths to reproducibility. *arXiv preprint arXiv:2504.07086*, 2025.

[22] Jian Hu. Reinforce++: A simple and efficient approach for aligning large language models. *arXiv preprint arXiv:2501.03262*, 2025.

[23] Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. Omnimed-vqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22170–22183, 2024.

[24] Fanding Huang, Guanbo Huang, Xiao Fan, Yi He, Xiao Liang, Xiao Chen, Qinting Jiang, Faisal Nadeem Khan, Jingyan Jiang, and Zhi Wang. Beyond the exploration-exploitation trade-off: A hidden state approach for llm reasoning in rlvr. *arXiv preprint arXiv:2509.23808*, 2025.

[25] 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.

[26] Yuxian Jiang, Yafu Li, Guanxu Chen, Dongrui Liu, Yu Cheng, and Jing Shao. Rethinking entropy regularization in large reasoning models. *arXiv preprint arXiv:2509.25133*, 2025.

[27] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018.

[28] Tianjian Li, Yiming Zhang, Ping Yu, Swarnadeep Saha, Daniel Khashabi, Jason Weston, Jack Lanchantin, and Tianlu Wang. Jointly reinforcing diversity and quality in language model generations. *arXiv preprint arXiv:2509.02534*, 2025.

[29] Yiwei Li, Peiwen Yuan, Shaoxiong Feng, Boyuan Pan, Bin Sun, Xinglin Wang, Heda Wang, and Kan Li. Turning dust into gold: Distilling complex reasoning capabilities from llms by leveraging negative data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18591–18599, 2024.

[30] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 1650–1654. IEEE, 2021.

[31] 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*, 2025.

[32] Miao Lu, Han Zhong, Tong Zhang, and Jose Blanchet. Distributionally robust reinforcement learning with interactive data collection: Fundamental hardness and near-optimal algorithms. *Advances in Neural Information Processing Systems*, 37:12528–12580, 2024.

[33] 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.

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

[35] Gongrui Nan, Siye Chen, Jing Huang, Mengyu Lu, Dexun Wang, Chunmei Xie, Weiqi Xiong, Xianzhou Zeng, Qixuan Zhou, Yadong Li, et al. Ngrpo: Negative-enhanced group relative policy optimization. *arXiv preprint arXiv:2509.18851*, 2025.

[36] 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.

[37] Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 337–347. Springer, 2025.

[38] Jinyoung Park, Jeehye Na, Jinyoung Kim, and Hyunwoo J Kim. Deepvideo-r1: Video reinforcement fine-tuning via difficulty-aware regressive grpo. *arXiv preprint arXiv:2506.07464*, 2025.

[39] 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.

[40] John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015.

[41] 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.

[42] Yi Shen, Jian Zhang, Jieyun Huang, Shuming Shi, Wenjing Zhang, Jiangze Yan, Ning Wang, Kai Wang, Zhaoxiang Liu, and Shiguo Lian. Dast: Difficulty-adaptive slow-thinking for large reasoning models. *arXiv preprint arXiv:2503.04472*, 2025.

[43] 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 *Proceedings of the Twentieth European Conference on Computer Systems*, pages 1279–1297, 2025.

[44] Christian Walder and Deep Tejas Karkhanis. Pass@k policy optimization: Solving harder reinforcement learning problems. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[45] Chen Wang, Lai Wei, Yanzhi Zhang, Chenyang Shao, Zedong Dan, Weiran Huang, Ge Lan, and Yue Wang. Targeted exploration via unified entropy control for reinforcement learning. *arXiv preprint arXiv:2604.14646*, 2026.

[46] 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. *Advances in Neural Information Processing Systems*, 37:95095–95169, 2024.

[47] Renxi Wang, Haonan Li, Xudong Han, Yixuan Zhang, and Timothy Baldwin. Learning from failure: Integrating negative examples when fine-tuning large language models as agents. *arXiv preprint arXiv:2402.11651*, 2024.

[48] Zengzhi Wang, Fan Zhou, Xuefeng Li, and Pengfei Liu. Octothinker: Mid-training incentivizes reinforcement learning scaling. *arXiv preprint arXiv:2506.20512*, 2025.

[49] Xumeng Wen, Zihan Liu, Shun Zheng, Shengyu Ye, Zhirong Wu, Yang Wang, Zhijian Xu, Xiao Liang, Junjie Li, Ziming Miao, et al. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base llms. *arXiv preprint arXiv:2506.14245*, 2025.

[50] An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement, 2024.

[51] Xinhao Yao, Lu Yu, Xiaolin Hu, Fengwei Teng, Qing Cui, Jun Zhou, and Yong Liu. The debate on rlvr reasoning capability boundary: Shrinkage, expansion, or both? a two-stage dynamic view. *arXiv preprint arXiv:2510.04028*, 2025.

[52] Edward Yeo, Yuxuan Tong, Xinyao Niu, Graham Neubig, and Xiang Yue. Demystifying long chain-of-thought reasoning in llms. In *ICLR 2025 Workshop on Deep Generative Model in Machine Learning: Theory, Principle and Efficacy*, 2025.

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

[54] 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.

[55] Enci Zhang, Xingang Yan, Wei Lin, Tianxiang Zhang, and Lu Qianchun. Learning like humans: Advancing llm reasoning capabilities via adaptive difficulty curriculum learning and expert-guided self-reformulation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6630–6644, 2025.

[56] Jixiao Zhang and Chunsheng Zuo. GRPO-LEAD: A difficulty-aware reinforcement learning approach for concise mathematical reasoning in language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5642–5665. Association for Computational Linguistics, November 2025.

[57] 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 *European Conference on Computer Vision*, pages 169–186. Springer, 2024.

[58] Shijie Zhang, Guohao Sun, Kevin Zhang, Xiang Guo, and Rujun Guo. Clpo: Curriculum learning meets policy optimization for llm reasoning. *arXiv preprint arXiv:2509.25004*, 2025.

[59] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023.

[60] Xingjian Zhang, Siwei Wen, Wenjun Wu, and Lei Huang. Edge-grpo: Entropy-driven grpo with guided error correction for advantage diversity. *arXiv preprint arXiv:2507.21848*, 2025.

[61] Zhanke Zhou, Xiangyu Lu, Chentao Cao, Brando Miranda, Tongliang Liu, Bo Han, and Sanmi Koyejo. Codapo: Confidence and difficulty-adaptive policy optimization for post-training language models. In *2nd AI for Math Workshop@ ICML 2025*, 2025.

[62] Xinyu Zhu, Mengzhou Xia, Zhepei Wei, Wei-Lin Chen, Danqi Chen, and Yu Meng. The surprising effectiveness of negative reinforcement in LLM reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

A Related Works

A.1 Advantage Estimation in Reinforcement Learning.

In Proximal Policy Optimization (PPO), advantage estimation relies on the Generalized Advantage Estimation (GAE [40]) framework. Nevertheless, the critic model required by GAE introduces non-trivial computational overhead. To mitigate this cost, GRPO [41] and REINFORCE++ [22] replace the critic with lightweight baselines (batch-averaged rewards) or group-relative rewards for advantage computation. Building on this, several methods have pointed out and addressed the limitations of GRPO's group relative advantage estimation. Dr.GRPO [31] enhances token efficiency by removing length and standard deviation normalization terms, while DAPO [53] balances the policy gradient loss at the token level to mitigate the insufficient gradient contribution of long sequences in long CoT scenarios. However, the advantage symmetry phenomenon inherent in GRPO's advantage estimation remains under-explored, with the existing literature lacking an in-depth analysis of this property.

A.2 GRPO Variants for Exploration Enhancement.

Recent studies have observed that RLVR may improve greedy accuracy while reducing solution diversity at larger sampling budgets [54]. To mitigate this issue, several methods [10, 18, 34, 14] explicitly encourage exploration through refined optimized objectives. Negative-reinforcement approaches [62, 35] suppress incorrect trajectories or reduce the relative reinforcement of correct ones, thereby redistributing probability mass toward alternative solutions. Pass@k-oriented methods [7, 44] instead directly optimize the collective utility of multiple sampled responses, encouraging diverse solutions.

From the perspective of advantage symmetry, these methods can be interpreted as different ways of breaking **group-level symmetry** to increase exploration. However, they primarily focus on how to strengthen exploration, while overlooking the training instability brought by the symmetry breaking. Our controlled experiments reveal that stronger group-level asymmetry is not always preferable: although down-weighting positive advantages substantially improves Pass@k, excessive or persistent negative dominance can cause training collapse. This finding motivates our attenuation strategy of A-GRAE, which promotes exploration when needed while preserving learning stability.

A.3 GRPO Variants for Difficulty Adaptation.

Another line of work improves GRPO by adapting the learning signal to problem difficulty. Since queries of different difficulty contribute unequally to policy optimization, difficulty-aware methods assign larger weights to challenging or informative samples [56, 42, 55, 2, 9], while curriculum-based approaches adjust the training distribution according to some designed data selection principles [61, 58]. These strategies generally improve overall accuracy by preventing optimization from being dominated by already-solved or less informative queries.

However, existing approaches generally impose a fixed difficulty preference or course paradigm, while how the optimal preference changes throughout training remains less explored. Our controlled experiments instead reveal a stage-dependent pattern: easier queries yield larger gains in the early stage, whereas harder queries become more beneficial as the model have reached a plateau in the latter stage. Accordingly, A-GRAE dynamically shifts its learning focus from easier to harder queries rather than maintaining a fixed difficulty bias.

B Supplementary Proof

B.1 Proof of Equation 4

Proof. We begin by reviewing the objective function of GRPO under the notation used in our formulation.

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim\mathcal{Q},\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q)} \left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\left\{ \min\left(\rho_{i,t}(\theta)A_{i,t}, \text{clip}(\rho_{i,t}(\theta), 1\pm\epsilon)A_{i,t}\right)\right\} - \beta\mathbb{D}_{\text{KL}}(\pi_\theta||\pi_{\text{ref}})\right], \tag{11}$$

where $\rho_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|q,o_{i,<t})}$, and β is the coefficient for the KL divergence term.

To better understand the model's learning dynamics under the verifiable reward setting and to derive the specific reweighting formulation, we omit the regularization components firstly (e.g., KL term & clipping operation) and focus on the core policy optimization objective:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim\mathcal{Q},\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q)} \left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\rho_{i,t}(\theta)A_{i,t}\right], \tag{12}$$

Taking the gradient with respect to θ:

$$\begin{aligned}
\nabla_\theta\mathcal{J}_{\text{GRPO}}(\theta) &= \mathbb{E}_{q\sim\mathcal{Q},\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q)} \left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}A_{i,t}\nabla_\theta\rho_{i,t}(\theta)\right] \\
&= \mathbb{E}_{q\sim\mathcal{Q},\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q)} \left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}A_{i,t}\frac{\nabla_\theta\pi_\theta(o_{i,t}\mid q,o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}\mid q,o_{i,<t})}\right] \\
&= \mathbb{E}_{q\sim\mathcal{Q},\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q)} \left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}A_{i,t}\frac{\pi_\theta(o_{i,t}\mid q,o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}\mid q,o_{i,<t})}\nabla_\theta\log\pi_\theta(o_{i,t}\mid q,o_{i,<t})\right] \\
&= \mathbb{E}_{q\sim\mathcal{Q},\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q)} \left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\underbrace{\rho_{i,t}(\theta)A_{i,t}}_{\text{weight}}\underbrace{\nabla_\theta\log\pi_\theta(o_{i,t}\mid q,o_{i,<t})}_{\text{SFT}}\right].
\end{aligned} \tag{13}$$

We complete the proof of Eq. (4). As the weight equal to 1, GRPO degrade to the standard SFT function, thus it can be formulated as a reweighting variant of Supervised Fine-Tuning.

Then we consider the full GRPO objective function including the clipping mechanism and the KL regularization term:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim\mathcal{Q},\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}} \left[\frac{1}{G}\sum_{i=1}^{G} \frac{1}{|o_i|}\sum_{t=1}^{|o_i|} \mathcal{L}_{i,t}^{\text{CLIP}}(\theta) - \beta\mathbb{D}_{\text{KL}}(\pi_\theta||\pi_{\text{ref}}) \right], \tag{14}$$

where the clipped surrogate objective is defined as:

$$\mathcal{L}_{i,t}^{\text{CLIP}}(\theta) = \min\left(\rho_{i,t}(\theta)A_{i,t}, \text{clip}(\rho_{i,t}(\theta), 1-\epsilon, 1+\epsilon)A_{i,t}\right). \tag{15}$$

Taking the gradient with respect to θ, we apply the chain rule. For the first term, we introduce an indicator function $\mathbb{I}_{i,t}$ to represent the clipping condition:

$$\nabla_\theta\mathcal{L}_{i,t}^{\text{CLIP}}(\theta) = \mathbb{I}_{i,t}\cdot A_{i,t}\nabla_\theta\rho_{i,t}(\theta) = \mathbb{I}_{i,t}\cdot\rho_{i,t}(\theta)A_{i,t}\nabla_\theta\log\pi_\theta(o_{i,t}\mid q, o_{i,<t}), \tag{16}$$

where $\mathbb{I}_{i,t} = 1$ if the ratio ρ is within the trust region or if the update further improves the objective without being clipped, and $\mathbb{I}_{i,t} = 0$ otherwise.

For the KL term, using the standard approximation $\nabla_\theta\mathbb{D}_{\text{KL}}(\pi_\theta||\pi_{\text{ref}}) \approx \nabla_\theta\log\pi_\theta$, the total gradient becomes:

$$\nabla_\theta\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|} \Big(\underbrace{\mathbb{I}_{i,t}\cdot\rho_{i,t}(\theta)A_{i,t}}_{\text{Gated Reward Weight}} - \underbrace{\beta\frac{\nabla_\theta\mathbb{D}_{\text{KL}}}{\nabla_\theta\log\pi_\theta}}_{\text{KL Penalty Weight}} \Big)\nabla_\theta\log\pi_\theta(o_{i,t}\mid q, o_{i,<t}) \right]. \tag{17}$$

Defining the effective weight as $W_{i,t} = \mathbb{I}_{i,t}\rho_{i,t}(\theta)A_{i,t} - \beta\psi_{i,t}$, we obtain:

$$\nabla_\theta\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|} W_{i,t}\cdot\nabla_\theta\log\pi_\theta(o_{i,t}\mid q, o_{i,<t}) \right]. \tag{18}$$

This confirms that GRPO, even with all regularization terms, maintains the form of a **Gated Reweighted SFT**, where the gating is controlled by clipping and the scale is adjusted by the KL penalty.

B.2 Proof of Equation 5

Given a group of sampled responses indexed by $i \in \{1, \ldots, G\}$, partitioned into a set of correct trajectories G_{pos} and incorrect trajectories G_{neg}. Let the advantage A_i be computed via GRAE as:

$$A_i = \frac{r_i - \mu}{\sigma}, \quad \text{where } \mu = \frac{1}{G}\sum_{j=1}^{G} r_j, \quad \sigma = \sqrt{\frac{1}{G}\sum_{j=1}^{G}(r_j-\mu)^2}. \tag{19}$$

Then, the sum of absolute advantages for correct trajectories equals that of incorrect trajectories:

$$\sum_{i\in G_{pos}} |A_i| = \sum_{i\in G_{neg}} |A_i|. \tag{20}$$

Proof. First, we examine the sum of the standardized advantages over the entire group. By definition:

$$\begin{aligned}
\sum_{i=1}^{G} A_i &= \sum_{i=1}^{G}\frac{r_i-\mu}{\sigma} \\
&= \frac{1}{\sigma}\left(\sum_{i=1}^{G} r_i - \sum_{i=1}^{G}\mu\right) \\
&= \frac{1}{\sigma}\left(\sum_{i=1}^{G} r_i - G\cdot\frac{1}{G}\sum_{j=1}^{G} r_j\right) \\
&= \frac{1}{\sigma}\left(\sum_{i=1}^{G} r_i - \sum_{i=1}^{G} r_i\right) = 0.
\end{aligned} \tag{21}$$

This confirms the zero-sum property of the standardized advantages. Now, we decompose the summation into the two disjoint sets G_{pos} and G_{neg} (assuming the group contains mixed results, i.e., $G_{pos} \cup G_{neg} = G$):

$$\sum_{i=1}^{G} A_i = \sum_{i \in G_{pos}} A_i + \sum_{i \in G_{neg}} A_i = 0. \tag{22}$$

Rearranging the terms, we obtain:

$$\sum_{i \in G_{pos}} A_i = - \sum_{i \in G_{neg}} A_i. \tag{23}$$

In the context of verifiable rewards (e.g., binary outcomes $r \in \{0, 1\}$), correct trajectories receive higher rewards than the group mean ($r_i > \mu$), implying $A_i > 0$ for $i \in G_{pos}$. Conversely, incorrect trajectories receive lower rewards ($r_i < \mu$), implying $A_i < 0$ for $i \in G_{neg}$. We have:

$$\sum_{i \in G_{pos}} |A_i| = \sum_{i \in G_{neg}} |A_i|. \tag{24}$$

Thus, the magnitude of policy updates attributed to correct trajectories is strictly equivalent to that of incorrect ones.

B.3 Proof of Theorem 1

In this section, we first derive Theorem 1 induced by the advantage term without clipping and KL regularization for clarity, and then extend the theorem to the full GRPO objective, confirming that it remains valid in a more general setting.

Theorem 1 (Behavior-Space Logit Update) *Assume that the set of all possible behaviors is $B = \{b_i\}_{i=1}^{N}$, consisting of sampled set G and unsampled set U. Given the intragroup advantage sum $C = \sum_{o_i \in G} A_i$, the behavior-level logit update of b_i under gradient ascent can be expressed as:*

$$\Delta h_{b_i} = \eta \left[\mathbb{I}(b_i \in G) A_{b_i} - C \pi_{b_i} \right]. \tag{25}$$

where η denotes the learning rate and π_{b_i} is the model's current sampling probability in behavior space for path b_i.

Proof. Fix a prompt x. Denote by $B = \{b_1, b_2, \ldots, b_N\}$ the behavior space, where each b_i represents a complete trajectory or response. For the current prompt, the sampled group is $G = \{b^{(1)}, b^{(2)}, \ldots, b^{(g)}\} \subseteq B$, with group size g, and the unsampled set is $U = B \backslash G$. In LLM reasoning, B is combinatorially large while g is typically small; hence U contains the overwhelming majority of possible trajectories.

Let $h = (h_{b_1}, h_{b_2}, \ldots, h_{b_N})^\top$ be the behavior-level logits. The induced policy over B is given by the Softmax parameterization

$$\pi_{b_i} = \pi_h(b_i \mid x) = \frac{\exp(h_{b_i})}{\sum_{j=1}^{N} \exp(h_{b_j})}. \tag{26}$$

Thus, h_{b_i} directly controls the probability mass assigned to trajectory b_i: increasing h_{b_i} raises π_{b_i}, while decreasing it lowers π_{b_i}. Here we first omit clipping and KL regularization, and consider the simplified policy-gradient objective as

$$J_{\text{pg}}(h) = \sum_{b_k \in G} \widehat{A}_{b_k} \log \pi_{b_k}, \tag{27}$$

where $\widehat{A}_{b_k}$ denotes the group-relative advantage of the sampled trajectory b_k.

For an arbitrary trajectory $b_i \in B$, the gradient of J_{pg} with respect to its logit h_{b_i} is

$$\frac{\partial J_{\text{pg}}}{\partial h_{b_i}} = \sum_{b_k \in G} \widehat{A}_{b_k} \frac{\partial \log \pi_{b_k}}{\partial h_{b_i}}. \tag{28}$$

By the standard log-softmax derivative,

$$\frac{\partial \log \pi_{b_k}}{\partial h_{b_i}} = \mathbb{I}(b_i = b_k) - \pi_{b_i}. \tag{29}$$

Substituting Eq. (29) into Eq. (28) yields

$$\begin{aligned}\frac{\partial J_{\text{pg}}}{\partial h_{b_i}} &= \sum_{b_k \in G} \widehat{A}_{b_k} \big(\mathbb{I}(b_i = b_k) - \pi_{b_i}\big) \\ &= \sum_{b_k \in G} \widehat{A}_{b_k} \mathbb{I}(b_i = b_k) - \pi_{b_i} \sum_{b_k \in G} \widehat{A}_{b_k}.\end{aligned} \tag{30}$$

Assuming the sampled trajectories are distinct, the first term is nonzero only when $b_i \in G$, and therefore

$$\sum_{b_k \in G} \widehat{A}_{b_k} \mathbb{I}(b_i = b_k) = \mathbb{I}(b_i \in G) \widehat{A}_{b_i}.$$

Together with the definition of C, we obtain

$$\frac{\partial J_{\text{pg}}}{\partial h_{b_i}} = \mathbb{I}(b_i \in G) \widehat{A}_{b_i} - C\pi_{b_i}. \tag{31}$$

Applying gradient ascent with learning rate $\eta > 0$, we can obtain:

$$\Delta h_{b_i} = \eta \left[\mathbb{I}(b_i \in G) \widehat{A}_{b_i} - C\pi_{b_i} \right]. \tag{32}$$

We complete the proof.

Extension of Theorem 1 considering the clipping and KL terms. We now refine Theorem 1 by incorporating the clipping mechanism and the KL regularization term in the complete GRPO objective. For a sampled trajectory $b_k \in G$, define the importance ratio $\rho_{b_k} = \frac{\pi_{b_k}}{\pi_{\text{old},b_k}}$. The clipped surrogate assigns an effective weight to each sampled trajectory:

$$w_{b_k} = \rho_{b_k} \widehat{A}_{b_k} \cdot \mathbb{I}\left[\big(\widehat{A}_{b_k} \geq 0 \ \wedge \ \rho_{b_k} \leq 1 + \epsilon\big) \vee \big(\widehat{A}_{b_k} < 0 \ \wedge \ \rho_{b_k} \geq 1 - \epsilon\big) \right]. \tag{33}$$

Here w_{b_k} denotes the actual advantage-driven gradient weight after importance reweighting and clipping. With this effective weight, the refined Theorem 1 under the complete GRPO objective are

$$\Delta h_{b_i} = \eta \left[\underbrace{\mathbb{I}(b_i \in G) w_{b_i} - \pi_{b_i} \sum_{b_k \in G} w_{b_k}}_{\text{advantage-driven term}} + \underbrace{\beta\big(\pi_{\text{ref},b_i} - \pi_{b_i}\big)}_{\text{KL-regularization term}} \right]. \tag{34}$$

Therefore, even in the complete formulation, if $b_i \in U$, the term $\mathbb{I}(b_i \in G) w_{b_i}$ vanishes, meaning unsampled correct trajectories still receive no direct positive reinforcement from the environment rewards. They are updated only by the normalization force (the summation of weights) and the KL penalty, which forces them to track the reference distribution rather than exploring novel high-reward regions.

Proof. The clipped advantage component of the complete GRPO objective can be written as

$$J_{\text{adv}}(h) = \sum_{b_k \in G} \min\left(\rho_{b_k} \widehat{A}_{b_k}, \text{clip}\left(\rho_{b_k}, 1 - \epsilon, 1 + \epsilon\right) \widehat{A}_{b_k} \right). \tag{35}$$

By the definition of w_{b_k}, its logit gradient takes the weighted log-likelihood form

$$\frac{\partial J_{\text{adv}}}{\partial h_{b_i}} = \sum_{b_k \in G} w_{b_k} \frac{\partial \log \pi_{b_k}}{\partial h_{b_i}}. \tag{36}$$

Using the log-Softmax derivative

$$\frac{\partial \log \pi_{b_k}}{\partial h_{b_i}} = \mathbb{I}(b_i = b_k) - \pi_{b_i},$$

we obtain

$$\frac{\partial J_{\mathrm{adv}}}{\partial h_{b_i}} = \sum_{b_k \in G} w_{b_k} \big(\mathbb{I}(b_i = b_k) - \pi_{b_i}\big)$$
$$= \sum_{b_k \in G} w_{b_k} \mathbb{I}(b_i = b_k) - \pi_{b_i} \sum_{b_k \in G} w_{b_k}. \tag{37}$$

Assuming sampled trajectories are distinct, the first term simplifies to

$$\sum_{b_k \in G} w_{b_k} \mathbb{I}(b_i = b_k) = \mathbb{I}(b_i \in G) w_{b_i}.$$

Therefore,

$$\frac{\partial J_{\mathrm{adv}}}{\partial h_{b_i}} = \mathbb{I}(b_i \in G) w_{b_i} - \pi_{b_i} \sum_{b_k \in G} w_{b_k}. \tag{38}$$

It remains to include the KL regularization. In GRPO, the KL penalty pulls the current policy toward the reference policy π_{ref}. At the behavior level, this regularization contributes the reference-pullback update

$$\Delta h_{b_i}^{\mathrm{KL}} = \eta\beta\big(\pi_{\mathrm{ref},b_i} - \pi_{b_i}\big). \tag{39}$$

Combining Eq. (38) with Eq. (39), and applying a gradient ascent step with learning rate η:

$$\Delta h_{b_i} = \eta \left[\mathbb{I}(b_i \in G) w_{b_i} - \pi_{b_i} \sum_{b_k \in G} w_{b_k} + \beta\big(\pi_{\mathrm{ref},b_i} - \pi_{b_i}\big) \right], \tag{40}$$

Thus the proof is completed.

Extension of Theorem 1 to Parameterized Neural Policies .Theorem 1 treats behavior-level logits as independent variables for analytical clarity. We now show how its interpretation extends to a nonlinear policy π_θ with shared parameters.

Ignoring clipping and KL regularization for clarity, a gradient-ascent step under the group-relative policy-gradient objective is

$$\Delta\theta = \eta \sum_{b_i \in G} A_i \nabla_\theta \log \pi_\theta(b_i \mid x). \tag{41}$$

For an unsampled trajectory $u \notin G$, a first-order Taylor expansion gives

$$\Delta \log \pi_\theta(u \mid x) \approx \nabla_\theta \log \pi_\theta(u \mid x)^\top \Delta\theta = \eta \sum_{b_i \in G} A_i K_\theta(u, b_i \mid x), \tag{42}$$

where

$$K_\theta(u, b_i \mid x) = \langle \nabla_\theta \log \pi_\theta(u \mid x), \nabla_\theta \log \pi_\theta(b_i \mid x) \rangle. \tag{43}$$

This expression shows that parameter sharing can indeed change the probability of an unsampled trajectory. However, such a change is mediated entirely by its gradient alignment with sampled trajectories; there is no reward term associated with u itself.

To further characterize this indirect update, decompose

$$K_\theta(u, b_i \mid x) = \bar{K}_u + \delta_i, \tag{44}$$

where $\bar{K}_u$ denotes the group-shared alignment component and δ_i the sample-specific deviation. Since standard GRAE satisfies $\sum_{b_i \in G} A_i = 0$, we obtain

$$\Delta \log \pi_\theta(u \mid x) \approx \eta \sum_{b_i \in G} A_i \delta_i. \tag{45}$$

Thus, the common component of parameter sharing cancels exactly.

Furthermore, assuming that, conditional on u being unsampled, the normalized advantage A_i and the sample-specific alignment deviation δ_i are uncorrelated,

$$\mathbb{E}[A_i \delta_i \mid x, u \notin G] = \mathrm{Cov}(A_i, \delta_i \mid x, u \notin G) = 0. \tag{46}$$

Consequently,

$$\mathbb{E}\left[\Delta \log \pi_\theta(u \mid x) \mid x, u \notin G\right] \approx 0. \tag{47}$$

Therefore, shared nonlinear parameters may induce incidental changes in unsampled trajectories, but standard GRAE provides no systematic reward-directed update toward an unseen trajectory before it is sampled. This is the parameter-space counterpart of the behavior-space result in Theorem 1.

B.4 Proof of Theorem 2

Theorem 2 (Update Magnitude with respect to Sample Difficulty). *Consider a group G of trajectories for one query (sample) with binary rewards $\{r_i\}_{i=1}^{G}$, the sum of absolute advantages over the group under Group Relative Advantage Estimation (GRAE) can be derived as:*

$$\sum_{i \in G} |A_i| = 2|G|\sqrt{p(1-p)}. \tag{48}$$

where $p = \sum_{i \in G} |r_i|/|G|$ denotes the empirical success probability of the corresponding sample within group G, and the proof can be found in Section B.4.

Proof. Given a group G of trajectories with binary rewards $r_i \in \{0, 1\}$, let $p = \sum_{i \in G} |r_i|/|G|$ denote the success probability. The group mean μ and standard deviation σ are:

$$\mu = p, \quad \sigma = \sqrt{p(1-p)} \tag{49}$$

The advantage in GRAE is defined as $A_i = (r_i - \mu)/\sigma$. Summing the absolute advantages over the group G, we partition the sum into successful ($r_i = 1$) and failed ($r_i = 0$) subsets:

$$\begin{aligned} \sum_{i \in G} |A_i| &= \sum_{r_i=1} \left| \frac{1-p}{\sqrt{p(1-p)}} \right| + \sum_{r_i=0} \left| \frac{-p}{\sqrt{p(1-p)}} \right| \\ &= (G \cdot p) \frac{1-p}{\sqrt{p(1-p)}} + (G \cdot (1-p)) \frac{p}{\sqrt{p(1-p)}} \\ &= \frac{2Gp(1-p)}{\sqrt{p(1-p)}} \\ &= 2|G|\sqrt{p(1-p)} \end{aligned} \tag{50}$$

This completes the proof, and a visualization of the relationship between update magnitude and sample difficulty is provided below, illustrating the derived curve:

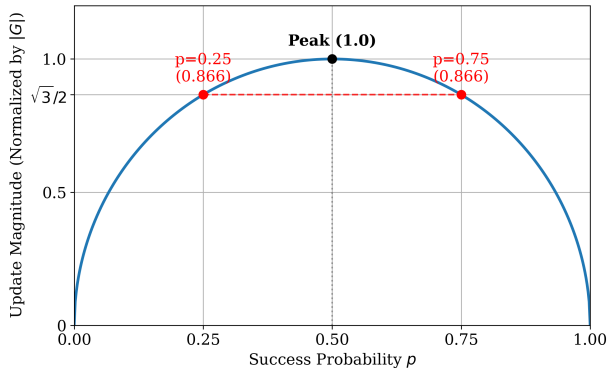

Figure 4: **Visualization of Update Magnitude with Respect to Sample Difficulty.** The update magnitude reaches its peak at $p = 0.5$, simple samples (e.g., $p = 0.75$) and hard samples (e.g., $p = 0.25$) exhibiting the same deviation from $p = 0.5$ are assigned identical importance weights.

B.5 The Logits Update of Negative-Dominant Group in Behavior Space

Theorem 1 (Behavior-Space Logit Update) *Assume that the set of all possible behaviors is $B = \{b_i\}_{i=1}^{N}$, consisting of sampled set G and unsampled set U. Given the intragroup advantage sum $C = \sum_{o_i \in G} A_i$, the behavior-level logit update of b_i under gradient ascent can be expressed as:*

$$\Delta h_{b_i} = \eta \left[\mathbb{I}(b_i \in G) A_{b_i} - C\pi_{b_i} \right]. \tag{51}$$

where η denotes the learning rate and π_{b_i} is the model's current sampling probability in behavior space for path b_i.

Down-weighting the advantages of correct paths leads to $C < 0$, and we subsequently examine the dynamics of probability updates across the following cases.

Case A. For the sampled positive responses $b_i \in G_{pos}$:

$$\Delta h_{b_i} = \eta(A_{pos} - C\pi_{b_i}), \tag{52}$$

where $A_{pos} > 0$, $(-C\pi_{b_i}) > 0$ thus $\Delta h_{b_i} > 0$. Hence, the sampling probability of a correct response will strictly increase.

Case B. For the sampled negative responses $b_i \in G_{neg}$:

$$\Delta h_{b_i} = \eta(A_{neg} - C\pi_{b_i}), \tag{53}$$

where $A_{neg} < 0$ but $(-C\pi_{b_i}) > 0$. Compared to GRPO($C = 0$) which directly penalizes the incorrect responses, the logits of overconfident negative responses can even be increased. As reinforcement learning sharpens the output distribution [18, 28, 32, 54], overconfident erroneous responses occasionally displace correct ones in later stages, leading to a sharp surge in fully unsolved samples.

Case C. For the unsampled response $b_i \in U$:

$$\Delta h_{b_i} = \eta(-C\pi_{b_i}) \tag{54}$$

It can be observed that the output probabilities of all unsampled responses exhibit an increase, whereas they remain constant under GRPO. This phenomenon demonstrates that the Negative-Dominant group provides an exploration incentive, potentially facilitating the discovery of correct yet previously unsampled trajectories.

C Implementation Details

C.1 Dataset Introduction

Mathematical Reasoning We first conduct experiments on three widely recognized math benchmarks designed to evaluate the logical reasoning capabilities of LLM:

- **MATH [20]:** A comprehensive dataset containing challenging competition-level mathematics problems across various subjects.
- **AMC23:** Problems derived from the 2023 American Mathematics Competitions, representing high-school-level competitive reasoning.
- **AIME 2025:** The latest problems from the American Invitational Mathematics Examination, designed to test advanced problem-solving skills and long-chain reasoning.

Multi-modal Mathematical Reasoning To validate the universality of our method in Vision-Language Models (VLMs), we extend our evaluation to three multi-modal mathematical reasoning datasets:

- **Geo3K [33]:** A dataset focused on geometry problem-solving that requires synchronized understanding of both visual diagrams and textual descriptions.
- **MathVerse [57]:** A comprehensive benchmark designed to evaluate visual mathematical reasoning across a wide array of subjects and visual patterns.
- **MathVision [46]:** A large-scale challenge consisting of diverse mathematical problems sourced from real-world competitions with complex visual contexts.

Medical Imaging Reasoning Finally, to verify whether the proposed method enhances the model's ability to perceive and reason over fine-grained image details, we utilize the **HuatuoGPT-Vision** [4] dataset. This benchmark is a curated and combined dataset from several publicly available medical VQA benchmarks, including VQA-RAD [27], SLAKE [30], PathVQA [19], OmniMedVQA [23], and PMC-VQA [59]. It is specifically designed to assess medical domain expertise and visual grounding capabilities in complex clinical scenarios. Specifically, the training set comprises 600 MRI image-question pairs, while the evaluation is conducted on a diverse test suite consisting of 300 MRI, 300 CT, and 300 X-ray pairs following [37].

C.2 Hyperparameter Settings

For all reinforcement learning experiments, responses were generated with a temperature of 1.0 and a maximum completion length of 2048 tokens. The learning rate is 1e-6. During evaluation, we used a generation temperature of 0.6, a top-p value of 0.95, and set the maximum new tokens to 4096. To ensure stable results, we perform 16 runs for AIME2025, and AMC23, and 4 runs for other benchmarks, reporting the average performance across the respective runs. Note that we did not use a dynamic sampling strategy in DAPO to ensure fair overhead. Our experiments are conducted over a single node with 8 NVIDIA H200 GPUs.

For experiments trained on the MATH dataset, we used the following system prompt to guide the model's reasoning process: "Please reason step by step, and put your final answer within \boxed{}." The training batch size is set to 1024 with total 20 epochs. For each prompt, we generated 8 responses within the group. The reward was based on binary accuracy, where a correct final answer yielded a reward of 1 and an incorrect one yielded 0.

For experiments trained on the Geo3K dataset, we used the following system prompt to guide the model's reasoning process: "This is a multiple-choice question. You FIRST think about the reasoning process as an internal monologue and then provide the final choice from the given options. The reasoning process MUST BE enclosed within tags." The training batch size is set to 512 with total 20 epochs. The final answer MUST BE put in \boxed{}." For each prompt, we generated 5 responses within the group. The reward was based on 0.1*acc-reward + 0.9 * format-reward. As for the medical dataset, the training batch size is set to 32 with total 10 epochs, and the other settings are the same as those on Geo3K.

Our proposed A-GRAE method is highly accessible, requiring only a single hyperparameter: the scaling parameter α in Eq. (10). In our experiments, α is set to 1 for the Math dataset and 0.5 for the Geo3K and Medical datasets. We also conduct the sensitivity analysis to α in Section E. Following prior works [21, 5, 7, 62], we set our primary evaluation metric as Pass@k. Pass@k estimates the probability of generating a correct solution within k attempts. Unlike greedy decoding, Pass@k provides a more reliable assessment of a model's potential and capability boundaries [21, 5, 7, 62]. We employ the standard unbiased estimator, which generates n responses ($n \geq k$) for each question q, counts the correct responses c, and computes the metric as follows:

$$\text{Pass@}k = \mathbb{E}_{q\sim\mathcal{Q}}\left[1 - \frac{\binom{n-c}{k}}{\binom{n}{k}}\right] \tag{55}$$

D Supplementary Experimental Results

D.1 Experimental Results of Control Experiment using Llama-3.2-3B-Instruct

It is observed that the empirical patterns on Llama-3.2-3B-Instruct remain largely consistent with those detailed in the main text. Specifically, down-weighting the positive advantages at the group level significantly enhances reasoning performance, particularly in terms of the Pass@k metric. Conversely, no universally optimal difficulty re-weighting strategy exists at the sample level. It is noteworthy, however, that when k reaches 256, all evaluated methods underperform relative to the base model. This suggests that the backbone model itself often determines the efficacy of reinforcement learning, a finding consistent with recent literature [52, 62, 48]. Nevertheless, the Negative-Dominant group remains the most effective approach for preserving the fundamental capabilities of the base model.

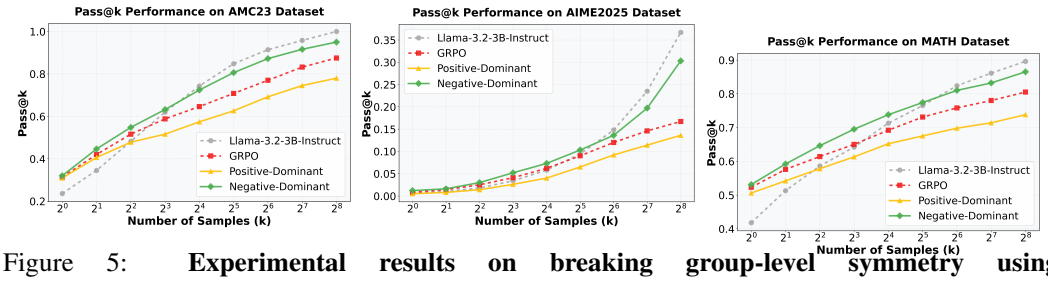

Figure 5: **Experimental results on breaking group-level symmetry using** `Llama-3.2-3B-Instruct`**.** The performance is evaluated using Pass@k ($k = \{1, 2, 4, 8, 16, 32, 64, 128, 256\}$) to provide a comprehensive analysis.

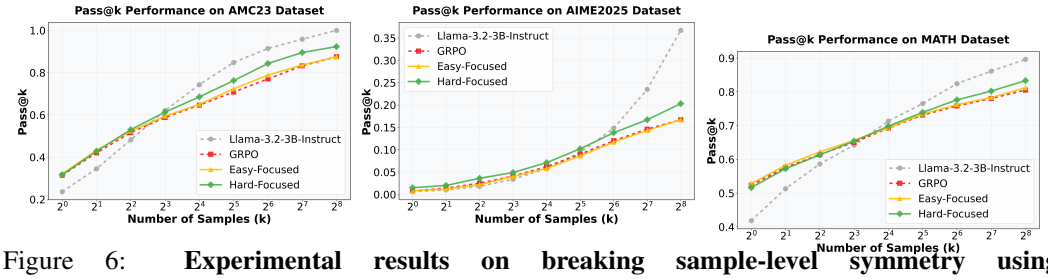

Figure 6: **Experimental results on breaking sample-level symmetry using** `Llama-3.2-3B-Instruct`**.** The performance is evaluated using Pass@k ($k = \{1, 2, 4, 8, 16, 32, 64, 128, 256\}$) to provide a comprehensive analysis.

D.2 Supplementary Experimental Results using DeepSeek-R1-7B

Table 4: Pass@k results on MATH, AIME 2025 and AMC23 with `DeepSeek-R1-7B`. **Bold** and underlined numbers denote the best and second-best results for each k. Specifically, We conduct 16 runs on AIME2025 and AMC23, and 4 runs on the remaining benchmarks.

Method	**Pass@k**									
k	1	2	4	8	16	32	64	128	256	**Avg.**
MATH										
Base Model	84.3 ±0.5	89.6 ±0.6	92.8 ±0.5	94.8 ±0.4	96.0 ±0.4	96.8 ±0.3	97.4 ±0.3	97.8 ±0.3	98.0 ±0.2	94.2
GRPO	92.4 ±0.3	95.3 ±0.4	96.5 ±0.4	97.0 ±0.3	97.3 ±0.3	97.7 ±0.2	98.0 ±0.2	98.3 ±0.2	98.6 ±0.2	96.8
GRPO + A-GRAE	92.9 ±0.2	95.7 ±0.3	96.9 ±0.3	97.5 ±0.3	97.9 ±0.3	98.3 ±0.2	98.6 ±0.2	**99.0** ±0.2	99.3 ±0.1	97.3
DAPO	92.6 ±0.3	95.3 ±0.4	96.4 ±0.4	97.2 ±0.3	97.5 ±0.3	97.8 ±0.3	98.0 ±0.2	98.2 ±0.2	98.6 ±0.2	96.8
DAPO + A-GRAE	93.0 ±0.2	95.6 ±0.3	97.0 ±0.3	97.5 ±0.3	98.1 ±0.2	98.2 ±0.2	98.4 ±0.2	98.7 ±0.2	99.3 ±0.1	97.3
Dr.GRPO	92.8 ±0.3	95.6 ±0.4	96.8 ±0.4	97.3 ±0.3	97.7 ±0.3	98.0 ±0.2	98.3 ±0.2	98.5 ±0.2	98.8 ±0.2	97.2
Dr.GRPO + A-GRAE	**93.3** ±0.2	**96.0** ±0.3	**97.3** ±0.3	**97.8** ±0.2	**98.3** ±0.2	**98.6** ±0.2	**98.8** ±0.2	**99.0** ±0.1	**99.4** ±0.1	**97.6**
AIME 2025										
Base Model	22.7 ±0.8	26.5 ±0.9	29.7 ±1.1	33.2 ±1.2	37.8 ±1.4	43.2 ±1.5	48.0 ±1.5	51.1 ±1.4	53.3 ±1.2	38.4
GRPO	28.5 ±0.5	35.0 ±0.7	41.4 ±0.9	47.7 ±1.1	53.1 ±1.2	57.6 ±1.4	62.1 ±1.3	67.2 ±1.3	73.3 ±1.1	51.8
GRPO + A-GRAE	29.6 ±0.4	36.0 ±0.6	42.6 ±0.8	48.9 ±0.9	54.0 ±1.1	58.3 ±1.2	63.0 ±1.2	68.3 ±1.1	74.3 ±1.0	52.8
DAPO	29.2 ±0.6	35.5 ±0.8	42.2 ±1.0	48.5 ±1.2	53.3 ±1.3	57.7 ±1.5	62.5 ±1.4	67.6 ±1.4	73.6 ±1.2	52.2
DAPO + A-GRAE	**30.3** ±0.4	**36.8** ±0.6	43.2 ±0.8	49.3 ±0.9	54.6 ±1.0	58.6 ±1.2	63.3 ±1.2	68.4 ±1.1	**75.0** ±0.9	**53.3**
Dr.GRPO	29.0 ±0.6	35.2 ±0.8	41.8 ±1.0	48.2 ±1.2	53.0 ±1.4	57.5 ±1.5	62.3 ±1.5	67.3 ±1.4	74.0 ±1.2	52.0
Dr.GRPO + A-GRAE	29.8 ±0.4	36.3 ±0.6	**43.3** ±0.8	**49.8** ±0.9	**55.2** ±1.1	**58.8** ±1.2	**63.5** ±1.2	**68.8** ±1.1	74.6 ±1.0	53.2
AMC23										
Base Model	64.2 ±0.6	73.0 ±0.7	80.1 ±0.8	85.5 ±0.9	89.8 ±0.9	92.8 ±0.8	95.2 ±0.7	97.4 ±0.5	**100.0** ±0.0	86.4
GRPO	82.8 ±0.4	89.3 ±0.5	92.3 ±0.6	94.0 ±0.7	95.2 ±0.7	96.1 ±0.6	97.2 ±0.5	98.8 ±0.4	**100.0** ±0.0	94.0
GRPO + A-GRAE	84.2 ±0.3	90.6 ±0.4	93.5 ±0.5	95.4 ±0.6	96.6 ±0.6	97.7 ±0.6	98.0 ±0.5	99.3 ±0.3	**100.0** ±0.0	95.0
DAPO	83.3 ±0.4	89.8 ±0.5	92.6 ±0.6	94.3 ±0.7	95.8 ±0.7	96.5 ±0.7	97.6 ±0.6	99.2 ±0.4	**100.0** ±0.0	94.3
DAPO + A-GRAE	**84.6** ±0.3	**91.6** ±0.4	**94.2** ±0.5	**96.3** ±0.6	97.0 ±0.6	97.8 ±0.6	98.2 ±0.5	**99.5** ±0.3	**100.0** ±0.0	**95.5**
Dr.GRPO	83.0 ±0.5	89.6 ±0.6	92.3 ±0.7	93.8 ±0.8	95.6 ±0.8	96.3 ±0.7	97.2 ±0.6	98.2 ±0.5	**100.0** ±0.0	94.0
Dr.GRPO + A-GRAE	84.5 ±0.3	91.3 ±0.4	93.6 ±0.5	96.0 ±0.6	**97.2** ±0.6	**98.0** ±0.6	**98.5** ±0.5	99.3 ±0.4	**100.0** ±0.0	95.4

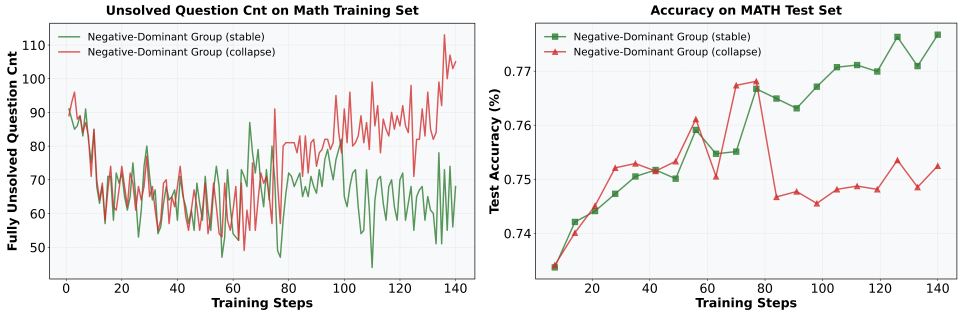

Figure 7: **Training collapse of Negative-Dominant group during training with `Qwen2.5-Math-7B`. Left Part:** The unsolved question count suddenly increase in the later stage. **Right Part:** The collapse leads to the significant performance decrease.

D.3 Training Collapse of Negative-Dominant Group

To further investigate the reliability of the Negative-Dominant configuration, we conducted multiple training runs under identical hyperparameter settings. As illustrated in Fig. 7, we observe a striking stochastic instability within this specific group. While one trial remains relatively stable throughout the training process, the other undergoes a catastrophic training collapse after approximately step 78. This divergence highlights that the Negative-Dominant approach is highly sensitive to initial conditions or sampling noise. In the "collapse" instance, the number of fully unsolved questions exhibits high-amplitude oscillations and a persistent upward trend, eventually peaking at over 110. This suggests that without proper regularization, the Negative-Dominant objective can easily lead the model into a divergent optimization path, where erroneous trajectories reinforce themselves and eventually overwhelm the model's reasoning capabilities. Such unpredictable behavior underscores the necessity of balancing negative trajectory weights to ensure consistent convergence.

D.4 Experimental Results in Further Analysis

Table 5: Ablation studies on MATH, AIME 2025 and AMC23 with `Qwen2.5-Math-7B`.

Method	**Pass@k**								
k	1	2	4	8	16	32	64	128	256
MATH									
Base Model	63.4	74.8	83.2	88.6	91.2	93.4	94.1	95.0	96.3
GRPO	76.5	82.3	86.1	88.8	90.3	92.6	93.5	93.9	95.0
GRPO + A-GRAE (sample level)	77.8	84.2	88.4	90.5	92.1	94.0	94.3	94.8	96.0
GRPO + A-GRAE (group level)	77.6	84.0	88.1	**91.2**	**93.0**	94.3	**95.3**	**95.6**	**97.0**
GRPO + A-GRAE (full)	**78.3**	**85.0**	**89.2**	91.0	92.5	**94.6**	95.0	95.5	96.5
AIME 2025									
Base Model	6.1	9.9	14.4	19.3	24.4	29.1	33.4	39.2	46.7
GRPO	10.3	14.3	18.7	23.1	27.5	31.8	36.1	40.8	46.7
GRPO + A-GRAE (sample level)	10.9	15.1	19.6	24.1	28.0	33.5	40.0	47.0	52.3
GRPO + A-GRAE (group level)	11.0	15.4	**20.3**	24.5	28.3	33.9	**41.0**	**49.6**	**60.0**
GRPO + A-GRAE (full)	**11.3**	**15.6**	19.8	**24.7**	**28.6**	**34.1**	39.2	47.8	56.7
AMC23									
Base Model	40.6	55.3	68.6	78.6	85.0	89.4	93.4	**97.3**	**100.0**
GRPO	59.2	66.7	72.1	76.4	80.6	84.8	88.3	90.8	92.5
GRPO + A-GRAE (sample level)	62.3	**70.4**	76.8	83.0	86.3	90.4	93.2	94.3	95.0
GRPO + A-GRAE (group level)	61.4	69.8	77.0	82.3	87.0	90.8	94.4	**97.3**	**100.0**
GRPO + A-GRAE (full)	**62.6**	70.0	**77.5**	**83.7**	**88.2**	**92.0**	**95.1**	96.8	97.5

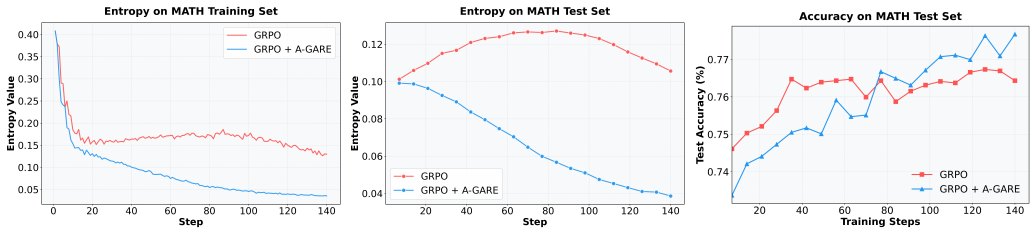

Figure 8: **Training dynamics using `Qwen2.5-Math-7B`. Left Part:** Model's entropy on the Math training set. **Center Part:** Model's entropy (actor entropy loss) on the Math test set. **Right Part:** The greedy decoding accuracy on the Math test set.

D.5 Comparison Experiments with Entropy Promoting Techniques

To verify whether the gains of our approach stem merely from a generic entropy-promoting effect, we conduct an additional entropy-matched control experiment. We implement an "Entropy-Promoting" baseline by tuning the `entropy_coeff` in the official *verl* framework. Specifically, by setting the coefficient to $8e{-}3$, we achieve a training-set entropy of 0.758 at step 105, which closely matches that of the Negative-Dominant group (0.742). As illustrated in Table 6, simply increasing the entropy does not improve reasoning performance. Despite explicitly encouraging higher entropy, the entropy-promoting baseline consistently underperforms vanilla GRPO across all k values. Upon further inspection, we observe that indiscriminate entropy promotion often leads to the "repetitive-output" phenomenon, where higher entropy corresponds to low-quality or redundant trajectories rather than meaningful diversity [26]. In contrast, our Negative-Dominant strategy, which breaks the symmetry in GRAE, consistently achieves the highest Pass@k scores, reaching 92.8% at $k = 16$ and maintaining superior scaling up to $k = 256$. This suggests that effective exploration must be encouraged fundamentally at the trajectory level through symmetry breaking, rather than through global entropy regularization.

Table 6: Comparison of MATH Pass@k (%) between Negative-Dominant group and entropy-promoting techniques.

Method	$k=1$	$k=2$	$k=4$	$k=8$	$k=16$	$k=256$
GRPO (Baseline)	76.5	82.3	86.1	88.8	90.3	92.4
Negative-Dominant	**77.2**	**83.8**	**87.6**	**90.4**	**92.8**	**94.5**
GRPO + Entropy-Promoting	75.8	80.3	83.6	86.5	88.4	89.0

D.6 Comprehensive Statistical Significance Analysis over Baselines

To assess A-GRAE's robustness across computational budgets, we conduct a fine-grained significance analysis using Welch's t-test across three primary baselines (GRPO, DAPO, and Dr.GRPO) at each interval $k \in [1, 256]$. As detailed in Table 7, A-GRAE consistently yields $p < 0.05$ across all benchmarks, confirming that its gains are statistically significant rather than stochastic fluctuations. Notably, on the MATH benchmark, Dr.GRPO + A-GRAE maintains $p < 0.005$ throughout scaling and retains a significant lead at $k = 256$ ($p = 0.0038$), demonstrating its efficacy in mitigating the capability boundary shrinkage observed in vanilla GRPO. For the high-variance AIME 2025 and AMC23 tasks, A-GRAE provides even more pronounced benefits; specifically, DAPO + A-GRAE achieves $p < 0.001$ across nearly all k intervals on AIME 2025. These consistent results across diverse baselines and datasets underscore the universal efficacy of breaking advantage symmetry to facilitate stable and efficient policy optimization during multi-attempt scaling.

E Sensitivity to Hyperparameter α

We evaluate the sensitivity of hyperparameter α using the Qwen2.5-Math-7B backbone under the GRPO training paradigm. As shown in Table 8, the results indicate that our method is fairly robust to α. The influence of α can be theoretically understood through the reformulated advantage estimation:

Table 7: Detailed p-values across scaling intervals ($k = 1$ to $k = 256$). p-values are calculated comparing the A-GRAE enhanced version against its respective vanilla baseline.

Comparison Pair	$k=1$	$k=2$	$k=4$	$k=8$	$k=16$	$k=32$	$k=64$	$k=128$	$k=256$
MATH Benchmark									
GRPO + A-GRAE vs. GRPO	.0032	.0018	.0025	.0041	.0084	.0115	.0128	.0155	.0121
DAPO + A-GRAE vs. DAPO	.0015	.0022	.0048	.0055	.0072	.0098	.0105	.0112	.0108
Dr.GRPO + A-GRAE vs. Dr.GRPO	.0024	.0009	.0011	.0022	.0015	.0028	.0035	.0042	.0038
AIME 2025 Benchmark									
GRPO + A-GRAE vs. GRPO	.0125	.0098	.0112	.0105	.0142	.0128	.0115	.0092	.0085
DAPO + A-GRAE vs. DAPO	.0016	.0012	.0008	.0005	.0008	.0006	.0004	.0005	.0003
Dr.GRPO + A-GRAE vs. Dr.GRPO	.0055	.0042	.0068	.0085	.0102	.0092	.0115	.0088	.0076
AMC23 Benchmark									
GRPO + A-GRAE vs. GRPO	.0045	.0038	.0042	.0051	.0048	.0062	.0075	.0082	.0065
DAPO + A-GRAE vs. DAPO	.0012	.0008	.0005	.0006	.0004	.0005	.0011	.0015	.0008
Dr.GRPO + A-GRAE vs. Dr.GRPO	.0008	.0012	.0009	.0007	.0011	.0015	.0008	.0022	.0015

Note: All reported p-values < 0.05 are considered statistically significant.

$A_i \cdot \min(1, w_s/\alpha)$. Here, α controls the degree of suppression on positive advantages: when α is small, the latter term always take 1, so the method behaves similarly to GRPO; as α increases, the positive advantage is gradually attenuated, and in the limit where $\alpha \to \infty$, it is nearly removed. Although excessively large α causes some performance degradation, learning remains effective overall, consistent with previous observations (GRPO-N [51]), which removes the positive advantage in GRPO entirely.

Table 8: Sensitivity analysis of α on Qwen2.5-Math-7B. Performance is measured by Pass@1 (%).

Method / Setting	MATH	AMC23	AIME2025
GRPO (Baseline)	76.5	59.2	10.3
$\alpha = 0.25$	77.8	61.8	10.9
$\alpha = 0.5$	77.8	61.8	11.0
$\alpha = 1.0$	**78.3**	62.6	**11.3**
$\alpha = 2.0$	78.0	**62.8**	**11.3**
$\alpha = 5.0$	76.8	62.3	10.8

F Limitations

Despite the theoretical insights and performance gains provided by A-GRAE, several limitations remain: While A-GRAE effectively mitigates the shrinkage of capability boundaries, it remains bounded by the inherent reasoning ceilings of the base model backbones, meaning it cannot fully compensate for fundamental knowledge gaps in smaller models. This study also primarily focuses on Reinforcement Learning with Verifiable Rewards (RLVR) using binary reward signals; therefore, the extent to which implicit advantage symmetry affects scenarios with dense, partial, or human-preference-based rewards remains to be explored in future work. Finally, although the method retains the efficiency of a critic-free architecture, finding the optimal balance between exploration and convergence in highly complex domains may still require extended training schedules.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly identify the "advantage symmetry" in GRPO as a bottleneck for exploration and difficulty adaptation. The paper subsequently provides theoretical proofs (Theorems 1 2) and experimental validation across seven benchmarks to support these claims.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: The paper includes a section on limitations in supplemental material, outlining the scope of the framework, and areas for future improvement.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The paper formalizes its findings through Theorem 1 (Logits Update) and Theorem 2 (Update Magnitude), with full mathematical proofs provided in the Supplementary Material (Sections B.3 and B.4).

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

Justification: The paper provides comprehensive details on datasets, experimental setups, and methodologies used, ensuring that the results can be reproduced accurately.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code is included in the supplemental material and will be open-sourced upon acceptance.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.

- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies relevant experimental details, including data splits, number of samples, and hyperparameters, ensuring transparency and reproducibility of the results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Error bars are not reported due to the high computational cost.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Detailed information on computational resources is provided in the supplementary materials to ensure reproducibility.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This research aligns with the NeurIPS Code of Ethics, ensuring responsible conduct throughout the study.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses potential societal impacts, with positive impacts mentioned in the introduction, and negative impacts mentioned in the limitations section (see supplement materials).

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper introduces a mathematical refinement to an optimization algorithm and does not release new large-scale pre-trained models or high-risk datasets.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All used datasets (e.g., MATH, Geo3k) and models (Qwen, Llama) are properly cited and credited to their original authors.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The proposed A-GRAE framework is mathematically defined in Section 5, and its integration into existing workflows is clearly documented.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The study does not involve human subjects or crowdsourcing; it is based on automated training and evaluation against public benchmarks.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No human subjects were involved in the research.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The LLM is utilized as the baseline modal, which is not the core innovation of this research. This is clarified in the method and experiment sections.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.