HGPO: Homogeneous-Aware Group-Relative Policy Optimization

Anonymous ACL submission

Abstract

Reinforcement Learning from Verifiable Rewards (RLVR) is an effective paradigm for improving the reasoning ability of large language models, but GRPO wastes a substantial fraction of training signal. When all sampled responses for a prompt receive the same reward, GRPO's group-normalized advantage collapses to zero and the prompt contributes no policy gradient. In this work, we propose Homogeneous-Aware Group-Relative Policy Optimization (HGPO), a minimal extension of GRPO that preserves standard normalization on mixed groups while assigning homogeneous-composition-aware advantages to homogeneous groups. The correction emphasizes the rarer homogeneous case in the current batch: rare all-correct groups receive stronger positive reinforcement, whereas rare all-incorrect groups receive stronger negative suppression. Experiments on Qwen3-1.7B, Qwen3-4B, and Qwen3-8B across six mathematical reasoning benchmarks show that HGPO delivers the best overall performance across all three backbones, with especially clear improvements on several competition-style tasks.

1 Introduction

In recent years, the reasoning capabilities of Large Language Models (LLMs) have garnered increasing attention (OpenAI, 2025; Guo et al., 2025; Comanici et al., 2025). Reinforcement Learning with Verifiable Rewards (RLVR) is a promising paradigm for improving such capabilities. By using rule-based outcome rewards, RLVR has shown particular effectiveness in domains with objectively checkable answers, such as mathematics and coding (Shao et al., 2024).

GRPO computes advantages by standardizing the rewards of responses sampled for the same prompt:

$$\hat{A}_{\text{GRPO}}(x, a_i) = \frac{r(a_i) - \bar{r}}{\sigma_r + \varepsilon}, \tag{1}$$

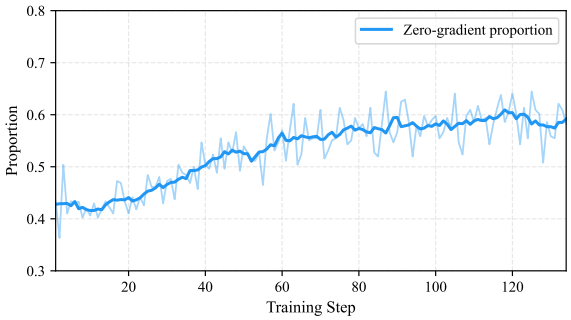


Figure 1: The proportion of prompts yielding zero gradients per batch during GRPO training on the DAPO-Math-17K dataset.

where x denotes the prompt and a_i denotes the i-th response sampled from $\pi_\theta(\cdot \mid x)$. Here, $\bar{r}$ and σ_r are the mean and standard deviation of the group rewards. This group-wise normalization highlights informative relative differences while suppressing noisy reward fluctuations, which is one reason GRPO is widely adopted in practice.

Despite these benefits, GRPO relies on a small fixed number of rollouts per prompt, which creates a vulnerability to signal collapse. When all sampled responses for a prompt receive identical rewards, either all correct or all incorrect, the group mean equals every individual reward. Consequently, the standardized advantage becomes zero, and the corresponding policy gradient vanishes. Such homogeneous-reward scenarios arise naturally during training: early on, difficult prompts often yield uniformly incorrect responses, whereas later, easy prompts may yield uniformly correct responses.

As shown in Figure 1, homogeneous groups are common in practice. Early in training, nearly half of the prompts in a batch can produce zero gradients under GRPO. Although this proportion decreases as the policy improves, it remains substantial throughout training, implying the waste of com-

pute resources.

Meanwhile, recent work (Zhu et al., 2026) decomposes the learning signal into reinforcing correct responses and penalizing incorrect responses, referred to as Positive Sample Reinforcement (PSR) and Negative Sample Reinforcement (NSR), respectively. Their weighted combination, W-REINFORCE, restores non-zero updates for homogeneous prompts by applying fixed-sign reinforcement or suppression. However, in our experiments, W-REINFORCE can become unstable and collapse into incoherent outputs when difficult prompts dominate training.

To address this issue, we propose Homogeneous-Aware Group-Relative Policy Optimization (HGPO). HGPO extends GRPO by assigning homogeneous-composition-aware advantages to homogeneous groups instead of always discarding them. The key idea is simple: the strength of these updates should depend on the current composition of homogeneous prompts. Rare all-correct groups should receive stronger positive reinforcement, while common all-incorrect groups should be penalized more weakly; when the model is already strong on a task, the same mechanism shifts toward stronger suppression of the remaining rare all-incorrect groups. In degenerate batches where only one homogeneous type appears, the correction for that common type can be zero, so HGPO should be interpreted as a rare-homogeneous-case correction rather than a guarantee of non-zero updates for every homogeneous prompt. HGPO implements this intuition while preserving the standard GRPO normalization on mixed groups.

Our contributions are summarized as follows:

- We identify a practical failure mode of GRPO: homogeneous-reward groups remain common enough during RLVR training to create substantial sample inefficiency.
- We propose Homogeneous-Aware Group-Relative Policy Optimization (HGPO), which augments GRPO with a homogeneous-composition-aware correction that recovers learning signal from rare homogeneous cases without changing the mixed-group objective.
- We derive the signed homogeneous-group gradient supplied by HGPO and bound the additional second moment introduced by the correction, clarifying how HGPO restores training signal while controlling variance.
- We validate the method on six mathematical reasoning benchmarks across Qwen3-1.7B, Qwen3-4B, and Qwen3-8B. HGPO achieves the best overall performance across all three backbones and substantially improves over strong RLVR baselines.

2 Preliminaries

Group Relative Policy Optimization (GRPO) (Shao et al., 2024) is a variant of Proximal Policy Optimization (PPO) (Schulman et al., 2017) tailored for LLM post-training. Compared to PPO, GRPO removes the value function and estimates advantages from a group of responses sampled for the same prompt. For a question-answer pair (q, a), the behavior policy $\pi_{\theta_{\text{old}}}$ samples G responses $\{o_i\}_{i=1}^{G}$. The advantage of the i-th response is then computed by normalizing the corresponding reward within the group:

$$\hat{A}_{i,t} = \frac{R_i - \text{mean}(\{R_i\}_{i=1}^{G})}{\text{std}(\{R_i\}_{i=1}^{G})}, \tag{2}$$

where $R_i = 1$ indicates a correct response and $R_i = 0$ indicates an incorrect one.

Similar to PPO, GRPO optimizes a clipped surrogate objective with a directly imposed KL penalty term:

$$\begin{aligned}\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{(q,a)\sim\mathcal{D},\,\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q)}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|} \\ \min\Big(r_{i,t}(\theta)\hat{A}_{i,t},\ \text{clip}\big(r_{i,t}(\theta), 1-\epsilon, 1+\epsilon\big)\hat{A}_{i,t}\Big) \\ -\beta\,\mathbb{D}_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}})\Bigg],\end{aligned} \tag{3}$$

where

$$r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})}. \tag{4}$$

Here, π_θ denotes the current policy, $\pi_{\theta_{\text{old}}}$ denotes the behavior policy used for sampling, and π_{ref} denotes the reference policy. The KL term constrains the updated policy from drifting too far from the reference model. The advantage $\hat{A}_{i,t}$ captures the intra-group reward ranking: responses with rewards above the group mean receive positive advantages, while those below the mean receive negative advantages.

For homogeneous groups with identical rewards, the advantage becomes zero as individual rewards match the mean. Although GRPO's normalization improves training stability, it renders these samples ineffective for gradient updates, thereby reducing

data utilization. This inefficiency becomes particularly pronounced for tasks that are either trivially easy or impossibly hard.

3 Motivation

A natural way to address GRPO's zero-gradient problem is to assign fixed advantages to homogeneous groups. Recent work (Zhu et al., 2026) shows that purely reinforcing correct responses or purely suppressing incorrect responses can produce useful learning signals, and their weighted combination yields W-REINFORCE. This suggests a simple remedy: assign a fixed positive advantage to homogeneous correct groups and a fixed negative advantage to homogeneous incorrect groups.

Yet this direct W-REINFORCE-style remedy is not sufficiently stable. As illustrated in Figure 2, W-REINFORCE can collapse under the same RLVR training setup used in our comparisons when difficult prompts dominate training, because uniformly applying fixed-sign updates to homogeneous groups ignores how informative those groups are at the current training stage. When correct responses are very rare, fully correct groups should be emphasized much more strongly than fully incorrect ones. Conversely, when the model already solves most prompts, the remaining fully incorrect groups become the more informative signal.

Therefore, the goal is not merely to make homogeneous groups update the policy, but to do so in a way that remains compatible with GRPO's stability. The desired correction should satisfy two properties: (1) it should preserve the standard relative normalization on mixed groups, where GRPO already works well; and (2) it should recover learning signal mainly from the rarer homogeneous case, while downweighting the homogeneous case that dominates the current batch. This observation naturally motivates our homogeneous-composition-aware formulation of HGPO.

4 Method

4.1 HGPO Objective and Adaptive Advantage

Homogeneous-Aware Group-Relative Policy Optimization (HGPO) keeps GRPO's mixed-group normalization but replaces the homogeneous zero-advantage rule with a homogeneous-composition-aware correction. The intuition is asymmetric but simple: among homogeneous prompts, if fully correct groups are rare, they are more informative and should receive stronger reinforcement; if fully incorrect groups are rare, they become the more informative failure signal and should receive stronger suppression. This yields a correction that recovers learning signal from rare homogeneous cases without changing the behavior of GRPO where it already works well.

HGPO keeps the clipped surrogate structure of GRPO and modifies only the advantage term. The objective is defined as:

$$\mathcal{J}_{\text{HGPO}}(\theta) = \mathbb{E}_{(q,a)\sim\mathcal{D},\,\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q)}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|} \min\Big(r_{i,t}(\theta)\hat{A}_{i,t},\ \text{clip}\big(r_{i,t}(\theta),1-\epsilon,1+\epsilon\big)\hat{A}_{i,t}\Big)\Bigg]. \tag{5}$$

The shaped advantage is defined as:

$$\hat{A}_{i,t} = \begin{cases} \dfrac{R_i - \mu_G}{\sigma_G + \epsilon}, & \text{if } 0 < \mu_G < 1, \\ \lambda\,(1-\rho_{\text{hom}}), & \text{if } \mu_G = 1, \\ -\lambda\,\rho_{\text{hom}}, & \text{if } \mu_G = 0, \end{cases} \tag{6}$$

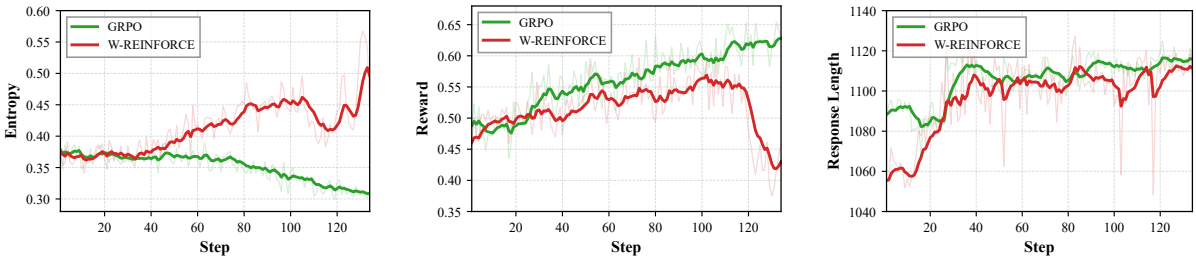
where $\mu_G = \text{mean}(\{R_i\}_{i=1}^{G})$ and $\sigma_G = \text{std}(\{R_i\}_{i=1}^{G})$ denote the mean and unbiased sample standard deviation of rewards within the group, ϵ is a small numerical constant, and λ is a scaling factor that controls the overall magnitude of the homogeneous-group correction. In our implementation, λ corresponds to the `factor` field in the algorithm configuration.

The adaptive term ρ_{hom} is computed from the homogeneous prompts in the current training batch. Let $\mathcal{H}^+$ denote the set of prompts whose sampled responses are all correct, and let $\mathcal{H}^-$ denote the set of prompts whose sampled responses are all incorrect:

$$\rho_{\text{hom}} = \frac{|\mathcal{H}^+|}{|\mathcal{H}^+| + |\mathcal{H}^-|}. \tag{7}$$

This ratio is used only when at least one homogeneous prompt appears in the batch; if all prompts are mixed, HGPO simply reduces to the standard GRPO advantage for that batch. Equivalently, the implementation first computes $z_1 = 1 - \rho_{\text{hom}}$ and $z_0 = -\rho_{\text{hom}}$, and then assigns λz_1 to all-correct groups and λz_0 to all-incorrect groups.

This adaptive mechanism rescales the correction according to the balance between the two homogeneous cases. When $\rho_{\text{hom}} \approx 0$, most homogeneous prompts are fully incorrect and fully correct groups are rare; HGPO therefore assigns strong



(a) Entropy (b) Reward (c) Response Length

Figure 2: RLVR training dynamics under GRPO and W-REINFORCE training paradigms. The fixed-sign homogeneous updates in W-REINFORCE can lead to training collapse. The figure compares a stable run (GRPO) with a collapsed run (W-REINFORCE). The collapse is characterized by a sharp decline in reward and a simultaneous surge in entropy.

positive advantages ($\approx \lambda$) to all-correct groups and mild negative advantages (≈ 0) to all-incorrect groups. Conversely, when $\rho_{\text{hom}} \approx 1$, fully incorrect groups are rare; HGPO assigns mild positive advantages (≈ 0) to all-correct groups and strong negative advantages ($\approx -\lambda$) to all-incorrect groups. At the extremes, if all homogeneous prompts in a batch are of one type, the correction for that common type is exactly zero. This behavior is intentional: HGPO does not force every homogeneous prompt to update the model, but instead concentrates homogeneous-group updates on the rarer case that is more informative under the current training state. In all regimes, the mixed-group advantage is left untouched.

4.2 Gradient and Variance Analysis

We now analyze how HGPO modifies the surrogate-gradient estimator compared to GRPO. The analysis is not a proof of global policy improvement; rather, it explains what training signal HGPO adds on prompts that GRPO would ignore, and why the additional stochastic variation is controlled by the homogeneous composition and the scale λ. We use the standard implementation convention of minimizing the negative surrogate and omit clipping and KL terms for clarity.

Directional signal on homogeneous groups. Let

$$\Psi_G(\theta) = \frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|} r_{i,t}(\theta)s_{i,t}, \qquad (8)$$
$$s_{i,t} = \nabla_\theta \log \pi_\theta(o_{i,t} \mid q, o_{i,<t}).$$

denote the average score-ratio gradient of a sampled group. For mixed groups, HGPO and GRPO use the same advantage and therefore produce the same surrogate-gradient term. For homogeneous groups, GRPO contributes zero, whereas HGPO contributes

$$g_{\text{hom}}^{\text{HGPO}} = \begin{cases} -\lambda(1-\rho_{\text{hom}})\Psi_G(\theta), & \mu_G = 1, \\ \lambda\rho_{\text{hom}}\Psi_G(\theta), & \mu_G = 0. \end{cases} \qquad (9)$$

Under gradient descent on the negative surrogate, the all-correct case moves parameters in the $+\Psi_G$ direction and therefore increases the likelihood of successful sampled trajectories, while the all-incorrect case moves in the $-\Psi_G$ direction and decreases the likelihood of failed trajectories. Thus HGPO supplies an explicit positive-reinforcement signal for all-correct homogeneous groups and an explicit suppression signal for all-incorrect homogeneous groups, exactly where GRPO discards the prompt-level update. The coefficients further emphasize the rarer homogeneous case: the all-correct signal is strongest when all-correct groups are rare, and the all-incorrect suppression is strongest when all-incorrect groups are rare.

Advantage-centering sanity check. HGPO preserves the same batch-level advantage centering as GRPO:

$$\sum_{q\in\mathcal{B}}\sum_{i=1}^{G}\hat{A}_i^{\text{HGPO}}(q) = \sum_{q\in\mathcal{B}}\sum_{i=1}^{G}\hat{A}_i^{\text{GRPO}}(q) = 0. \qquad (10)$$

The equality holds because mixed groups are centered within each prompt, while the homogeneous correction is centered across the batch by ρ_{hom}.

Variance control. The same adaptive coefficients also control the extra stochastic variation introduced by HGPO. Let $p_+ = \Pr(\mu_G = 1)$, $p_- = \Pr(\mu_G = 0)$, and $h = p_+ + p_-$ denote the homogeneous-group rate. In the population

analogue where $h > 0$ and $\rho_{\text{hom}} = p_+/h$, and assuming $\mathbb{E}[\|\Psi_G\|^2 \mid \mu_G \in \{0,1\}] \leq C$, the additional second moment induced by homogeneous groups is bounded by

$$\begin{aligned}\Delta M_2 &\leq \lambda^2 C\big[p_+(1-\rho_{\text{hom}})^2 + p_-\rho_{\text{hom}}^2\big] \\ &= \lambda^2 C \frac{p_+ p_-}{p_+ + p_-} \leq \frac{\lambda^2 C}{4}(p_+ + p_-).\end{aligned} \tag{11}$$

Because covariance is bounded by the corresponding second moment, this also controls the within-type variance contribution. The bound shows two stabilizing effects. First, the added stochastic variation is quadratic in the chosen scale λ, which is kept below the smallest non-zero GRPO majority-class scale in Appendix B. Second, the adaptive factors suppress the homogeneous type that dominates the batch: the bound is maximized when all-correct and all-incorrect homogeneous groups are balanced, and it shrinks when one type dominates. A fixed homogeneous bonus/penalty would instead scale with $\lambda^2 C(p_+ + p_-)$ and would not receive this composition-dependent reduction.

5 Experiments and Results

5.1 Experimental Setup

Training setup. We use Qwen3-1.7B, Qwen3-4B, and Qwen3-8B as the base models. For training, we use the DAPO-Math-17K corpus (Yu et al., 2026) and implement all RLVR methods with the verl framework (Sheng et al., 2025). The global training batch size is 256, with 8 rollouts generated per prompt. During training, the rollout temperature is set to 1.0, top-p is set to 1.0, and top-k is disabled. The maximum prompt length and response length are set to 2,048 and 4,096, respectively. We use a PPO mini-batch size of 256 and a learning rate of 1×10^{-6}. More hyperparameter settings can be found in Appendix C.

Baseline settings. We compare HGPO against GRPO, NGRPO (Nan et al., 2025), DAPO (Yu et al., 2026), and W-REINFORCE (Zhu et al., 2026). We also report the performance of the base models for reference.

Evaluation setup. We evaluate our models on six mathematical reasoning benchmarks: MATH500 (Lightman et al., 2024), AMC23 (MAA, 2023), AIME24 and AIME25 (MAA, 2025), OlympiadBench (He et al., 2024), and Minerva Math (Lewkowycz et al., 2022). During evaluation, we sample 16 responses per prompt with a temperature of 0.6 and a top-p of 0.95; the prompt templates are provided in Appendix D.2. We use Avg@16 as the primary evaluation metric.

5.2 Main Results

As shown in Table 1, we evaluate HGPO and the primary comparison baselines on six mathematical reasoning benchmarks using Avg@16 as the main metric. HGPO gives the strongest overall performance across all three backbones.

Performance on Qwen3-1.7B. On Qwen3-1.7B, HGPO achieves the best overall score of 44.9, improving over GRPO by +2.3 points, NGRPO by +7.7 points, DAPO by +4.7 points, and W-REINFORCE by +7.6 points. HGPO obtains the best result on five of the six benchmarks, including AIME24, AIME25, AMC23, MATH500, and Olympiad Bench. The only exception is Minerva Math, where GRPO is higher by 1.0 point. These results indicate that the adaptive homogeneous-group correction is especially useful when the smaller model frequently encounters uniformly solved or uniformly failed groups.

Performance on Qwen3-4B. On Qwen3-4B, HGPO shows the clearest gains. It reaches an overall score of 61.1, outperforming GRPO by +4.5 points, NGRPO by +10.1 points, DAPO by +8.4 points, and W-REINFORCE by +11.4 points. HGPO is also the best method on every individual benchmark. The improvements over GRPO are particularly large on AIME25 (+9.4), AIME24 (+8.3), and AMC23 (+3.9), showing that recovering learning signal from homogeneous groups can translate into broad gains beyond the training-domain math benchmarks.

Comparison with GRPO. Compared with GRPO, HGPO improves the overall Avg@16 by +2.3 on Qwen3-1.7B, +4.5 on Qwen3-4B, and +2.8 on Qwen3-8B. On Qwen3-8B, HGPO reaches the best overall score of 59.8, outperforming NGRPO by +7.3 points. HGPO obtains the best result on five of the six benchmarks, with especially clear improvements over GRPO on AIME24 (+4.4), AIME25 (+5.2), AMC23 (+4.6), and Olympiad Bench (+1.8). The only exception is Minerva Math, where GRPO is slightly higher by 0.1 point. These results show that the homogeneous-group correction remains useful at larger scale rather than only benefiting smaller and medium backbones.

Table 1: Avg@16 results (%) on six mathematical reasoning benchmarks. Overall is the average across all benchmarks. **Bold** denotes the best result and underline denotes the second best within each backbone group.

Method	AIME24	AIME25	AMC23	MATH500	Olympiad	Minerva	Overall
Qwen3-1.7B Backbone	14.2	11.3	43.8	72.6	39.4	29.0	35.0
W-REINFORCE	15.6	12.5	46.4	75.3	42.0	31.7	37.3
NGRPO	16.5	12.9	47.3	74.5	41.3	30.6	37.2
DAPO	19.4	14.2	50.9	77.9	45.2	33.6	40.2
GRPO	19.4	16.7	57.2	79.4	47.2	**35.6**	42.6
HGPO (ours)	**21.7**	**20.8**	**62.8**	**80.7**	**49.1**	34.6	**44.9**
Qwen3-4B Backbone	21.9	16.5	66.3	83.6	50.5	42.6	46.9
W-REINFORCE	23.3	21.0	70.3	85.6	54.1	44.0	49.7
NGRPO	28.3	24.0	70.3	86.0	54.5	42.7	51.0
DAPO	31.3	25.8	72.5	87.1	55.1	44.3	52.7
GRPO	38.8	27.9	80.6	89.2	58.0	45.4	56.6
HGPO (ours)	**47.1**	**37.3**	**84.5**	**90.4**	**60.7**	**46.5**	**61.1**
Qwen3-8B Backbone	26.7	19.6	67.3	83.8	52.7	42.1	48.7
W-REINFORCE	27.5	21.5	69.2	85.2	55.4	43.0	50.3
NGRPO	30.4	24.6	73.4	86.7	56.3	43.6	52.5
DAPO	32.9	24.8	74.8	88.3	58.3	44.3	53.9
GRPO	39.8	27.7	79.2	90.0	60.5	**44.9**	57.0
HGPO (ours)	**44.2**	**32.9**	**83.8**	**91.0**	**62.3**	44.8	**59.8**

Overall Performance. Taken together, the results show that HGPO is a strong and stable extension of GRPO. It consistently improves over the base model, NGRPO, DAPO, W-REINFORCE, and GRPO across all three backbones, and delivers the best overall performance in every backbone group. These gains are obtained without changing GRPO's mixed-group objective, isolating the value of recovering useful updates from homogeneous groups that GRPO would otherwise ignore.

5.3 Inference-time Scaling

To further examine whether the gains persist beyond the default Avg@16 setting, we plot Pass@k curves for $k \in \{1, 2, 4, 8, 16, 32, 64, 128\}$ on the competition-style mathematical reasoning benchmarks. As shown in Figures 3 to 5, HGPO consistently maintains strong scaling as the sampling budget increases across Qwen3-1.7B, Qwen3-4B, and Qwen3-8B. The advantage is especially clear on AIME-style benchmarks, where recovering signal from homogeneous groups improves both single-sample accuracy and high-budget Pass@k performance.

5.4 Ablation Studies

Fixed vs. adaptive homogeneous advantages. To isolate the contribution of the batch-adaptive homogeneous weighting, we compare HGPO with a fixed-advantage variant, HGPO-Constant. This variant keeps the same treatment of mixed groups as HGPO, but assigns a constant $+\lambda$ advantage to all-correct homogeneous groups and a constant $-\lambda$ advantage to all-incorrect homogeneous groups, without using the batch composition ratio ρ_{hom}. For HGPO-Constant, we set $\lambda = 0.3$, the same default scale used by HGPO, so the comparison isolates the effect of adaptive versus fixed homogeneous weighting.

As shown in Table 2, simply restoring non-zero updates for homogeneous groups is not sufficient to match HGPO. The adaptive version improves the overall Avg@16 over HGPO-Constant by $+5.0$, $+4.9$, and $+2.5$ points on Qwen3-1.7B, Qwen3-4B, and Qwen3-8B, respectively, and it is better on every benchmark. This confirms that the gains come not only from using additional homogeneous prompts, but also from weighting all-correct and all-incorrect homogeneous groups according to the current batch composition.

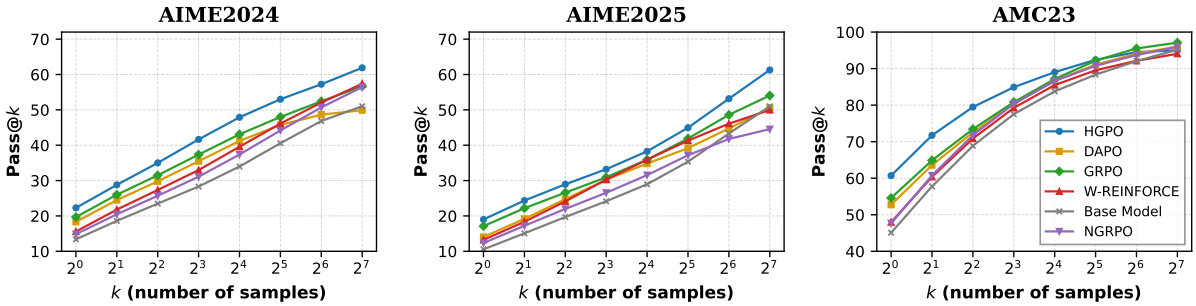


Figure 3: Pass@k curves of `Qwen3-1.7B` trained with different RLVR algorithms.

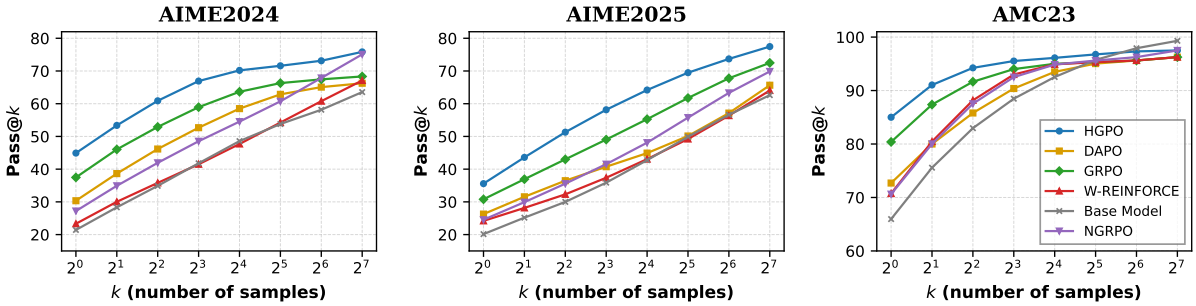


Figure 4: Pass@k curves of `Qwen3-4B` trained with different RLVR algorithms.

Impact of λ on Performance. To study the effect of the homogeneous-group correction magnitude, we analyze the global scaling factor λ. The full ablation results are summarized in Table 3. Specifically, we evaluate $\lambda \in \{0.1, 0.2, 0.3, 0.5, 0.8, 1\}$ on Qwen3-4B across six mathematical reasoning benchmarks using Avg@16 as the evaluation metric. The scaling factor has a measurable but not brittle effect on performance. A small non-zero value, $\lambda = 0.1$, reaches an overall Avg@16 of 59.2. Increasing the scale to $\lambda = 0.2$, 0.8, or 1 improves the average to 60.5, while $\lambda = 0.3$ gives the best overall score of 61.1. This indicates that homogeneous-group correction is beneficial across a broad range, but the default scale provides the strongest aggregate result. Combining these observations with the scale-consistency argument in Appendix B, we adopt $\lambda = 0.3$ as the default setting. This value stays below the unbiased-standard-deviation scale $1/\sqrt{G} \approx 0.354$ for $G = 8$, obtains the best overall score, and delivers strong results on the AIME-style benchmarks while remaining competitive on the remaining tasks. All main results in Table 1 therefore use $\lambda = 0.3$.

6 Related Work

Reinforcement Learning with Verifiable Rewards. Reinforcement Learning with Verifiable Rewards (RLVR) has become a practical paradigm for improving LLM reasoning in domains where outcomes can be checked automatically, such as mathematics, coding, and question answering (Muennighoff et al., 2025; Guo et al., 2025; Shao et al., 2024). Existing RLVR methods therefore tend to discard such zero-variance prompts either implicitly (Shao et al., 2024; Liu et al., 2025) or explicitly through filtering and dynamic sampling (Yu et al., 2026; Zheng et al., 2026). This is inefficient because rollout generation is a major bottleneck in online RL training (Zhong et al., 2025; Noukhovitch et al., 2025), and homogeneous groups can appear frequently during mathematical reasoning training. Our work addresses this specific inefficiency by recovering controlled learning signals from homogeneous groups while preserving the standard GRPO update on mixed groups.

Addressing signal loss in GRPO. Homogeneous reward groups cause signal loss in GRPO because their within-group rewards have zero variance, making the normalized advantage vanish. Prior work (Zhu et al., 2026) assign fixed advantages to correct and incorrect samples, respectively. However, such fixed corrections are insensitive to the current composition of groups. Other methods avoid or reshape the issue by filtering uniform-reward prompts, as in DAPO-style dynamic sam-

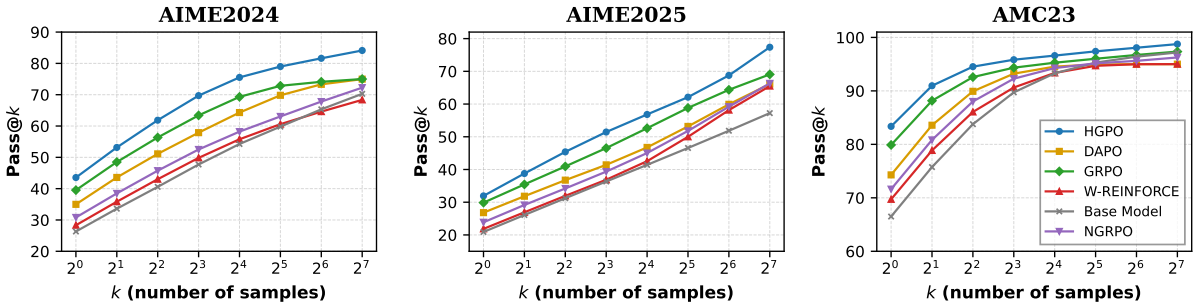


Figure 5: Pass@k curves of `Qwen3-8B` trained with different RLVR algorithms.

Table 2: Ablation of the batch-adaptive homogeneous weighting. HGPO-Constant replaces the adaptive homogeneous advantages with fixed $+\lambda/-\lambda$ coefficients while keeping the same mixed-group objective. Results are Avg@16 (%).

Backbone	Method	AIME24	AIME25	AMC23	MATH500	Olympiad	Minerva	Overall
Qwen3-1.7B	HGPO-Constant	17.7	14.8	51.6	77.2	45.0	33.2	39.9
	HGPO	**21.7**	**20.8**	**62.8**	**80.7**	**49.1**	**34.6**	**44.9**
Qwen3-4B	HGPO-Constant	35.8	32.1	78.4	88.8	57.2	45.1	56.2
	HGPO	**47.1**	**37.3**	**84.5**	**90.4**	**60.7**	**46.5**	**61.1**
Qwen3-8B	HGPO-Constant	40.6	30.0	79.2	89.4	60.0	44.5	57.3
	HGPO	**44.2**	**32.9**	**83.8**	**91.0**	**62.3**	**44.8**	**59.8**

Table 3: Ablation on the scaling factor λ (Qwen3-4B, Avg@16). **Bold** denotes the best and underline the second best per benchmark.

λ	AIME24	AIME25	AMC23	MATH500	Olympiad	Minerva	Avg.
0.1	41.5	32.7	84.5	90.5	60.0	45.8	59.2
0.2	45.2	34.2	**86.7**	90.4	60.6	45.8	60.5
0.3	**47.1**	37.3	84.5	90.4	60.7	**46.5**	**61.1**
0.5	43.4	33.3	85.8	90.6	60.7	46.1	60.0
0.8	41.9	**37.5**	85.8	90.5	**61.4**	46.1	60.5
1	44.0	35.4	86.1	**90.9**	60.3	46.4	60.5

pling (Yu et al., 2026), or by introducing additional reward construction, as in NGRPO (Nan et al., 2025). In contrast, HGPO preserves GRPO's standard advantage for mixed groups and intervenes only on zero-variance groups. Its batch-composition-aware correction amplifies the rarer all-correct or all-incorrect case while downweighting the homogeneous type that dominates the current batch.

7 Conclusion

In this work, we presented HGPO, a simple extension of GRPO that addresses the zero-gradient problem caused by homogeneous reward groups. Instead of assigning a fixed advantage to such groups, HGPO uses a homogeneous-composition-aware adaptive correction: all-correct groups receive stronger reinforcement when correct solutions are rare, while all-incorrect groups receive stronger suppression when failures are rare. This design improves data utilization while preserving the stability benefits of GRPO on mixed groups. Across mathematical reasoning benchmarks, HGPO achieves the best overall Avg@16 performance across all evaluated backbones.

Limitations

This work focuses primarily on mathematical reasoning benchmarks and is evaluated on Qwen3-1.7B, Qwen3-4B, and Qwen3-8B. Broader validation across other model families remains future work. Although HGPO removes the need for a task-specific fixed advantage, it still depends on the global scaling factor λ and on the estimation of the homogeneous-group composition during training. In extreme batches where homogeneous prompts are all of one type, the correction for that common type can become zero by design, so HGPO should be understood as emphasizing rare homogeneous cases rather than guaranteeing non-zero updates for every homogeneous prompt. Our theoretical analysis is also diagnostic: it characterizes the estimator under a stated independence condition, while the empirical results test whether the recovered homogeneous-group signal is useful in actual RLVR training. Understanding how these quantities transfer across model scales and training regimes remains future work.

References

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, and 1 others. 2024. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3828–3850.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, and 1 others. 2022. Solving quantitative reasoning problems with language models. *Advances in neural information processing systems*, 35:3843–3857.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's verify step by step. In *International Conference on Learning Representations*, volume 2024, pages 39578–39601.

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

MAA. 2023. American mathematics competitions - amc. https://maa.org/.

MAA. 2025. American invitational mathematics examination - aime. https://maa.org/.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori B Hashimoto. 2025. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20286–20332.

Gongrui Nan, Siye Chen, Jing Huang, Mengyu Lu, Dexun Wang, Chunmei Xie, Weiqi Xiong, Xianzhou Zeng, Qixuan Zhou, Yadong Li, and 1 others. 2025. Ngrpo: Negative-enhanced group relative policy optimization. *arXiv preprint arXiv:2509.18851*.

Michael Noukhovitch, Shengyi Huang, Sophie Xhonneux, Arian Hosseini, Rishabh Agarwal, and Aaron Courville. 2025. Asynchronous rlhf: Faster and more efficient off-policy rl for language models. In *International Conference on Learning Representations*, volume 2025, pages 4003–4029.

OpenAI. 2025. Introducing openai o3 and o4-mini.

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2025. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pages 1279–1297.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, and 1 others. 2026. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244.

Zhiyuan Zeng, Qinyuan Cheng, Zhangyue Yin, Yunhua Zhou, and Xipeng Qiu. 2025. Revisiting the test-time scaling of o1-like models: Do they truly possess test-time scaling capabilities? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4651–4665.

Haizhong Zheng, Yang Zhou, Brian Bartoldson, Bhavya Kailkhura, Fan Lai, Jiawei Zhao, and Beidi Chen. 2026. Act only when it pays: Efficient reinforcement learning for llm reasoning via selective rollouts. *Advances in Neural Information Processing Systems*, 38:124321–124346.

Yinmin Zhong, Zili Zhang, Bingyang Wu, Shengyu Liu, Yukun Chen, Changyi Wan, Hanpeng Hu, Lei Xia, Ranchen Ming, Yibo Zhu, and 1 others. 2025. Optimizing {RLHF} training for large language models with stage fusion. In *22nd USENIX Symposium on Networked Systems Design and Implementation (NSDI 25)*, pages 489–503.

Xinyu Zhu, Mengzhou Xia, Zhepei Wei, Wei-Lin Chen, Danqi Chen, and Yu Meng. 2026. The surprising effectiveness of negative reinforcement in llm reasoning. *Advances in Neural Information Processing Systems*, 38:126546–126573.

A Policy Gradient Derivation of HGPO

We analyze HGPO as a surrogate-gradient estimator relative to GRPO. For a query q, the old policy samples a group of k trajectories $G = \{\tau_1, \ldots, \tau_k\}$, with $\tau_i = (a_{i,1}, \ldots, a_{i,|\tau_i|})$ and $s_{i,t} = \nabla_\theta \log \pi_\theta(a_{i,t} \mid q, a_{i,<t})$. Let $R(q, \tau_i) \in \{0, 1\}$ denote the verifiable outcome reward. GRPO constructs the group statistics

$$\begin{aligned} \mu_G &= \frac{1}{k} \sum_{j=1}^{k} R(q, \tau_j), \\ \sigma_G &= \mathrm{std}\Big(\{R(q, \tau_j)\}_{j=1}^{k} \Big), \end{aligned} \tag{12}$$

and the group-centered surrogate advantage

$$\tilde{A}_i = R(q, \tau_i) - \mu_G. \tag{13}$$

Because μ_G is computed from the sampled group, including the current response, this is a GRPO surrogate estimator rather than an independent-baseline transformation of the exact policy gradient. GRPO employs the normalized advantage $\hat{A}_i^{\mathrm{GRPO}} = \tilde{A}_i / (\sigma_G + \epsilon)$ when $\sigma_G > 0$, where ϵ is used for numerical stability. In homogeneous groups, $\tilde{A}_i = 0$ for all i, causing the surrogate gradient to vanish. HGPO modifies this zero-gradient case by replacing the homogeneous-group advantage with homogeneous-composition-aware values:

$$\hat{A}_i^{\mathrm{HGPO}} = \begin{cases} \dfrac{R(q, \tau_i) - \mu_G}{\sigma_G + \epsilon}, & 0 < \mu_G < 1, \\ \lambda\,(1 - \rho_{\mathrm{hom}}), & \mu_G = 1, \\ -\lambda\,\rho_{\mathrm{hom}}, & \mu_G = 0, \end{cases} \tag{14}$$

where ρ_{hom} is the homogeneous solved ratio and λ is a scaling factor. During each policy update, ρ_{hom} is computed from sampled rewards and treated as a fixed coefficient; gradients are not propagated through this batch statistic. The homogeneous coefficient is not guaranteed to be non-zero for every homogeneous prompt: if $\rho_{\mathrm{hom}} = 0$, all-incorrect groups receive zero correction, and if $\rho_{\mathrm{hom}} = 1$, all-correct groups receive zero correction. This boundary behavior downweights the homogeneous type that fully dominates the batch. For readability, the subsequent theoretical derivations omit ϵ in mixed groups, while the implementation uses $\sigma_G + \epsilon$ as in the main text.

We use the probability ratio

$$r_{i,t}(\theta) = \frac{\pi_\theta(a_{i,t} \mid q, a_{i,<t})}{\pi_{\theta_{\mathrm{old}}}(a_{i,t} \mid q, a_{i,<t})}. \tag{15}$$

Assumption A.1 (Surrogate-gradient convention)**.** Trajectories in a sampled group are treated as Monte Carlo samples from the rollout policy. When differentiating the surrogate objective, the advantage values $\hat{A}_i$ and the batch statistic ρ_{hom} are treated as fixed coefficients. Clipping and KL regularization are omitted in the derivations below unless explicitly stated.

Assumption A.2 (Finite second moments)**.** The score-function averages used in the stochastic gradient estimators have bounded second moments. For variance comparisons, we use the scalar variance $\mathrm{Var}(g) = \mathbb{E}[\|g - \mathbb{E}g\|^2]$ and the corresponding covariance matrix $\mathrm{Cov}(g)$.

Assumption A.3 (Null advantage-gradient independence)**.** For each algorithm $\mathrm{alg} \in \{\mathrm{GRPO}, \mathrm{HGPO}\}$ and each sampled trajectory-token pair (i, t), the advantage coefficient $\hat{A}_i^{\mathrm{alg}}$ is independent of the score-ratio gradient term $r_{i,t}(\theta) s_{i,t}$.

All expectation and variance derivations in the following analysis are taken under Theorem A.3. This assumption is a null diagnostic condition, not a claim that real RLVR training satisfies independence. In practice, the usefulness of any reward-derived advantage depends on correlation between

rewards, sampled trajectories, and policy-gradient directions. The analysis below therefore characterizes what HGPO changes when that correlation is removed; the empirical sections evaluate whether the added homogeneous-group coefficients are useful in actual training. We write the gradient relations at the expectation level.

A.1 Expected Gradient Difference and Equality

For algorithm $\mathrm{alg} \in \{\mathrm{GRPO}, \mathrm{HGPO}\}$, define the expected surrogate gradient as

$$\bar{g}^{\mathrm{alg}}(\theta) = -\mathbb{E}_{q,G}\left[\tfrac{1}{k}\sum_{i=1}^{k}\tfrac{1}{|\tau_i|}\sum_{t=1}^{|\tau_i|} r_{i,t}(\theta)\hat{A}_i^{\mathrm{alg}} s_{i,t}\right]. \tag{16}$$

The expected gradient difference is therefore

$$\begin{aligned}\Delta\bar{g}(\theta) &= \bar{g}^{\mathrm{HGPO}}(\theta) - \bar{g}^{\mathrm{GRPO}}(\theta)\\ &= -\frac{1}{k}\sum_{i=1}^{k}\mathbb{E}_{q,G}\left[\frac{1}{|\tau_i|}\sum_{t=1}^{|\tau_i|} r_{i,t}(\theta)\left(\hat{A}_i^{\mathrm{HGPO}} - \hat{A}_i^{\mathrm{GRPO}}\right) s_{i,t}\right].\end{aligned} \tag{17}$$

Since $\hat{A}_i^{\mathrm{HGPO}} = \hat{A}_i^{\mathrm{GRPO}}$ when $0 < \mu_G < 1$, only homogeneous groups contribute to the expected difference:

$$\begin{aligned}\Delta\bar{g}(\theta) = &-\mathbb{E}_{q,G}\left[\mathbf{1}[\mu_G = 1]\frac{\lambda(1-\rho_{\mathrm{hom}})}{k}\sum_{i=1}^{k}\frac{1}{|\tau_i|}\sum_{t=1}^{|\tau_i|} r_{i,t}(\theta)s_{i,t}\right]\\ &+\mathbb{E}_{q,G}\left[\mathbf{1}[\mu_G = 0]\frac{\lambda\rho_{\mathrm{hom}}}{k}\sum_{i=1}^{k}\frac{1}{|\tau_i|}\sum_{t=1}^{|\tau_i|} r_{i,t}(\theta)s_{i,t}\right].\end{aligned} \tag{18}$$

Thus, at the estimator level, HGPO can be viewed as GRPO plus a homogeneous-composition-aware correction term whose contribution is zero on mixed groups and automatically scaled by the balance between all-correct and all-incorrect homogeneous groups. The correction can also be zero for the homogeneous type that fully dominates a batch.

Under Theorem A.3, this correction does not change the expected stochastic gradient. For either algorithm,

$$\begin{aligned}\bar{g}^{\mathrm{alg}}(\theta) &= -\mathbb{E}_{q,G}\left[\frac{1}{k}\sum_{i=1}^{k}\frac{1}{|\tau_i|}\sum_{t=1}^{|\tau_i|}\hat{A}_i^{\mathrm{alg}} r_{i,t}(\theta)s_{i,t}\right]\\ &= -\frac{1}{k}\sum_{i=1}^{k}\frac{1}{|\tau_i|}\sum_{t=1}^{|\tau_i|}\mathbb{E}\left[\hat{A}_i^{\mathrm{alg}} r_{i,t}(\theta)s_{i,t}\right]\\ &= -\frac{1}{k}\sum_{i=1}^{k}\frac{1}{|\tau_i|}\sum_{t=1}^{|\tau_i|}\mathbb{E}\left[\hat{A}_i^{\mathrm{alg}}\right]\mathbb{E}[r_{i,t}(\theta)s_{i,t}]\\ &= 0,\end{aligned} \tag{19}$$

where the second equality uses Theorem A.3 and the final equality follows from the score-function identity under importance weighting. Therefore,

$$\bar{g}^{\mathrm{GRPO}}(\theta) = \bar{g}^{\mathrm{HGPO}}(\theta) = 0. \tag{20}$$

Thus HGPO changes the homogeneous-group term inside the stochastic estimator but preserves the expected stochastic gradient under the null advantage-gradient independence condition. This result should not be read as a proof that HGPO improves the true policy-gradient direction; rather, it shows that the adaptive correction does not introduce a directional bias when the recovered homogeneous-group coefficients contain no directional information.

A.2 Gradient Variance Analysis of GRPO vs. HGPO

We compare the variance of stochastic policy-gradient estimators under GRPO and HGPO. Let Z^{alg} denote the random estimator whose expectation gives $\bar{g}^{\mathrm{alg}}(\theta)$ in Equation (16). Let $H \in \{\mathrm{mix}, +, -\}$ denote group type: $H = \mathrm{mix}$ if $0 < \mu_G < 1$, $H = +$ if $\mu_G = 1$, and $H = -$ if $\mu_G = 0$. Let $p_h = \Pr(H = h)$. Since the previous subsection shows that both expected gradients are zero under Theorem A.3, the variance comparison isolates the stochastic effect introduced by HGPO's homogeneous-group correction under the same null condition.

By the law of total covariance,

$$\begin{aligned}\mathrm{Cov}(Z^{\mathrm{alg}}) = &\sum_h p_h \,\mathrm{Cov}(Z_h^{\mathrm{alg}})\\ &+\sum_h p_h\left(m_h^{\mathrm{alg}} - m^{\mathrm{alg}}\right)\left(m_h^{\mathrm{alg}} - m^{\mathrm{alg}}\right)^{\top},\end{aligned} \tag{21}$$

where $Z_h^{\mathrm{alg}} = Z^{\mathrm{alg}} \mid H = h$, $m_h^{\mathrm{alg}} = \mathbb{E}[Z_h^{\mathrm{alg}}]$, and $m^{\mathrm{alg}} = \mathbb{E}[Z^{\mathrm{alg}}]$.

For GRPO, homogeneous groups contribute zero estimator value: $Z_+^{\mathrm{GRPO}} = Z_-^{\mathrm{GRPO}} = 0$ and $Z_{\mathrm{mix}}^{\mathrm{GRPO}} = Z_{\mathrm{mix}}^{\mathrm{HGPO}}$. Define

$$\Psi(q,G) = \tfrac{1}{k}\sum_{i=1}^{k}\tfrac{1}{|\tau_i|}\sum_{t=1}^{|\tau_i|} r_{i,t}(\theta)\, s_{i,t}, \tag{22}$$

and let $\bar{\psi}_h = \Psi(q,G) \mid H = h$. For HGPO, we have $Z_+^{\mathrm{HGPO}} = -\lambda\,(1-\rho_{\mathrm{hom}})\,\bar{\psi}_+$ and $Z_-^{\mathrm{HGPO}} = +\lambda\,\rho_{\mathrm{hom}}\,\bar{\psi}_-$.

The covariance difference decomposes as

$$\begin{aligned}&\mathrm{Cov}(Z^{\mathrm{HGPO}}) - \mathrm{Cov}(Z^{\mathrm{GRPO}})\\ &= \lambda^2\left(p_+\,(1-\rho_{\mathrm{hom}})^2\,\mathrm{Cov}(\bar{\psi}_+) + p_-\,\rho_{\mathrm{hom}}^2\,\mathrm{Cov}(\bar{\psi}_-)\right)\\ &\quad + \left(\mathrm{BetweenVar}_{\mathrm{HGPO}} - \mathrm{BetweenVar}_{\mathrm{GRPO}}\right),\end{aligned} \tag{23}$$

where BetweenVar denotes the second term in Equation (21). The adaptive scaling keeps the variance contribution context-sensitive: when $\rho_{\text{hom}} \approx 0$, the all-correct term $(1-\rho_{\text{hom}})^2 \approx 1$ is large but fully correct homogeneous groups are rare; conversely, when $\rho_{\text{hom}} \approx 1$, the all-incorrect term $\rho_{\text{hom}}^2 \approx 1$ is large but fully incorrect homogeneous groups are rare. This natural balancing mechanism helps limit variance inflation.

Interpretation under prompt difficulty. Under a simplified population view where each sampled response for a prompt is independently correct with probability p, homogeneous groups occur with probability $p^G + (1-p)^G$. Thus, HGPO differs most from GRPO when p is close to 0 or 1, where homogeneous groups are frequent. In the intermediate regime, both p^G and $(1-p)^G$ are small, so HGPO behaves similarly to GRPO and introduces limited additional variance.

Table 4: Policy update direction under each HGPO case.

Case	$\hat{A}_i^{\text{HGPO}}$	Policy update	Effect
Mixed, $R_i > \mu_G$	> 0	Increase	Reinforce
Mixed, $R_i < \mu_G$	< 0	Decrease	Suppress
All correct ($\mu_G = 1$)	$+\lambda\,(1-\rho_{\text{hom}})$	Increase if < 1	Adaptive exploit
All incorrect ($\mu_G = 0$)	$-\lambda\,\rho_{\text{hom}}$	Decrease if > 0	Adaptive suppress

B Understanding the Advantage Scale

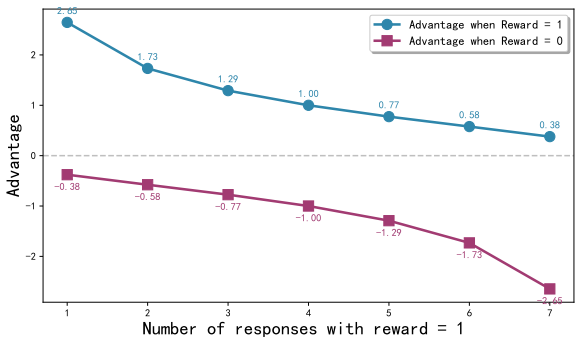


Figure 6: Advantage values for positive (Reward=1) and negative (Reward=0) samples as the number of positive samples varies within a group of 8 responses. The advantage for a correct response is maximized when correct responses are rare, illustrating how normalization amplifies rare high-reward events while suppressing common outcomes.

In GRPO, the advantage $\hat{A}_{i,t}$ is computed by standardizing rewards within each prompt-specific group, as shown in Equation (2). This normalization is useful because it emphasizes relative differences: minority outcomes within a group receive larger-magnitude advantages, whereas majority outcomes receive smaller ones.

As illustrated in Figure 6, when 7 responses are correct and 1 is incorrect, the normalized advantages under unbiased standard deviation are

$$A_{\text{correct}} \approx 0.354, \quad A_{\text{incorrect}} \approx -2.475.$$

Conversely, when only 1 response is correct and 7 are incorrect, the advantages become

$$A_{\text{correct}} \approx 2.475, \quad A_{\text{incorrect}} \approx -0.354.$$

This reweighting effect allows GRPO to amplify rare but informative outcomes while suppressing common ones, which is why HGPO leaves the mixed-group advantage in Equation 6 unchanged.

However, the same normalization degenerates when all sampled responses for a prompt receive the same reward. In that case, the group variance is zero, every advantage collapses to zero, and the entire prompt contributes no policy gradient. HGPO reduces this gap by assigning composition-aware advantages to homogeneous groups, while preserving GRPO's relative signal whenever a group contains both correct and incorrect responses. When only one homogeneous type appears in a batch, the correction for that type can be zero; in this case HGPO intentionally behaves like GRPO for that homogeneous type rather than forcing a fixed-sign update.

The scaling factor λ does not replace the adaptive homogeneous-composition term; it controls how aggressively HGPO exploits the recovered homogeneous-group signal. Choosing λ therefore balances two concerns.

Lower bound. Setting $\lambda = 0$ disables the homogeneous-group correction entirely and reduces HGPO to standard GRPO. Small λ values leave most of the zero-gradient prompts unexploited and provide only marginal improvements over GRPO.

Upper bound. When λ is too large, homogeneous-group updates can dominate the relative learning signal from mixed groups. Under the unbiased sample standard deviation used in our implementation, the smallest non-zero majority-class advantage magnitude produced by GRPO on a non-homogeneous group of size G is

Table 5: Comprehensive Hyperparameter Configuration for HGPO Math Training. The table details model, optimization, generation, and infrastructure settings derived from the training script

Category	Hyperparameter	Math training
Model/data	Base model	Qwen3-1.7B, 4B, and 8B
	Epochs	2
	Global batch size	256
	Max prompt length	2048
	Max response length	4096
	Truncation mode	Error
Rollout	Rollouts per prompt	8
	Temperature	1.0
	Top-p	1.0
	Top-k	Disabled (-1)
	Thinking template	False
	Accelerator	NVIDIA A100
Optimization	Learning rate	1×10^{-6}
	Warmup steps	10
	Weight decay	0.1
	PPO mini-batch size	256
	KL loss coefficient	0.0
	Entropy coefficient	0.0
	Gradient checkpointing	True
	Dynamic batch size	True
	Remove padding	True
Infrastructure	Tensor parallel size	1
	GPUs per node	8
	GPU memory utilization	0.65

$1/\sqrt{G}$, attained when exactly one response disagrees with the rest. For $G = 8$, this gives a scale of $1/\sqrt{8} \approx 0.354$, matching the majority-class value in Figure 6. Choosing λ below this scale helps keep the homogeneous-group correction consistent with GRPO's own signal magnitude.

C Detailed Experimental Setup

All experimental settings are summarized in Table 5.

D Detailed Descriptions of Methods

This section provides the objectives and implementation details for the baselines used in our experiments.

D.1 Objectives and Training Hyperparameters

The objectives of GRPO, DAPO, NGRPO, and W-REINFORCE are as follows:

$$\mathcal{L}_{\text{GRPO}}(\theta) = -\frac{1}{G}\sum_{i=1}^{G}\frac{1}{T}\sum_{t=1}^{T}\Bigg(\min\Big(r_{i,t}(\theta)\,A_{i,t},\ \text{clip}\big(r_{i,t}(\theta),\,1-\epsilon,\,1+\epsilon\big)\,A_{i,t}\Big) - \beta\,\mathbb{D}_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}})\Bigg), \tag{24}$$

where $r_{i,t}(\theta) = \dfrac{\pi_\theta(y_{i,t} \mid x,\, y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t} \mid x,\, y_{i,<t})}$.

$$\mathcal{J}_{\text{DAPO}}(\theta) = \mathbb{E}_{(q,a)\sim\mathcal{D},\,\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q)}\left[\frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}\min\Big(r_{i,t}(\theta)\,\hat{A}_{i,t},\ \text{clip}\big(r_{i,t}(\theta),\,1-\epsilon_{\text{low}},\,1+\epsilon_{\text{high}}\big)\,\hat{A}_{i,t}\Big)\right],$$
$$\text{s.t.}\quad 0 < \Big|\{o_i \mid \texttt{is_equivalent}(a, o_i)\}\Big| < G, \tag{25}$$

where

$$r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})}, \quad \hat{A}_{i,t} = \frac{R_i - \text{mean}(\{R_i\}_{i=1}^{G})}{\text{std}(\{R_i\}_{i=1}^{G})}. \tag{26}$$

DAPO maximizes the clipped surrogate in Equation (25); in a loss-minimization implementation, the optimized loss is its negative. The constraint keeps only prompts whose sampled responses contain both correct and incorrect answers, matching DAPO's dynamic sampling rule.

For NGRPO, let $\mathcal{R} = \{r_1, \ldots, r_G\}$ be the set of rewards from the G sampled responses. NGRPO augments it with a virtual maximum reward, $\mathcal{R}' = \mathcal{R} \cup \{r_{\max}\}$, and computes the advantage as

$$A'_i = \frac{r_i - \mu'_{\mathcal{R}}}{\sigma'_{\mathcal{R}} + \epsilon_{\text{std}}}, \quad \mu'_{\mathcal{R}} = \frac{1}{G+1}\left(\sum_{j=1}^{G} r_j + r_{\max}\right), \quad \sigma'_{\mathcal{R}} = \sqrt{\frac{1}{G+1}\sum_{r\in\mathcal{R}'}(r-\mu'_{\mathcal{R}})^2}. \tag{27}$$

The NGRPO objective is then defined as

$$\mathcal{J}_{\text{NGRPO}}(\theta) = \mathbb{E}_{q\sim P(Q),\,\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q)}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\min\Big(r_{i,t}(\theta)A'_i,\ \text{clip}\big(r_{i,t}(\theta),\,1-\epsilon,\,1+\epsilon\big)A'_i\Big)\right], \tag{28}$$

$$\mathcal{L}_{\text{PSR}}(\theta) = -\frac{1}{T}\sum_{t=1}^{T}\min\Big(r_t(\theta),\ \text{clip}\big(r_t(\theta),\,1-\epsilon,\,1+\epsilon\big)\Big),$$
$$\text{where } r_t(\theta) = \frac{\pi_\theta(y_t \mid x,\, y_{<t})}{\pi_{\theta_{\text{old}}}(y_t \mid x,\, y_{<t})}, \quad r(x, y) = 1, \tag{29}$$

Table 6: Prompt template for `Qwen3 series`.

```
<|im_start|>user
{input}
Please reason step by step, and put your final answer within \boxed{}.<|im_end|>
<|im_start|>assistant
```

$$\mathcal{L}_{\text{NSR}}(\theta) = -\frac{1}{T}\sum_{t=1}^{T} \min\Big(-r_t(\theta),\, -\text{clip}\big(r_t(\theta),\, 1-\epsilon,\, 1+\epsilon\big)\Big),$$
$$\text{where } r_t(\theta) = \frac{\pi_\theta(y_t \mid x,\, y_{<t})}{\pi_{\theta_{\text{old}}}(y_t \mid x,\, y_{<t})}, \quad r(x,y) = -1, \tag{30}$$

$$\mathcal{L}_{\text{W-REINFORCE}}(\theta) = \lambda\, \mathcal{L}_{\text{PSR}}(\theta) + \mathcal{L}_{\text{NSR}}(\theta), \tag{31}$$

where ϵ is the clip ratio and β is the KL penalty coefficient.

D.2 Prompt Templates

We adopt the prompt template for Qwen models following Zeng et al. (2025), as shown in Table 6.

E Use of AI Assistants

We utilized AI assistants to help with language editing and writing refinement. All technical content, experimental results, and scientific claims were verified by the authors.

F Artifacts Statements

F.1 Model Artifacts

The models utilized in our work, **Qwen3-1.7B, Qwen3-4B, and Qwen3-8B**, are released under the Apache License, allowing both research and commercial usage.

F.2 Data Artifacts

We employ publicly available benchmarks for evaluation, including MATH-500, OlympiadBench, AIME 24/25, AMC 23, and Minerva. These datasets are widely accessible in the open-source community. We utilize them strictly for non-commercial research purposes and ensure that our usage complies with the respective licenses and terms of use.