UPAO: UNBOUNDED POSITIVE ASYMMETRIC OPTIMIZATION FOR BREAKING THE EXPLORATION–STABILITY DILEMMA IN RL

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning (RL) has become the standard paradigm for enhancing the complex logical reasoning capabilities of large language models (LLMs). Modern RL frameworks improve sample efficiency by employing importance sampling (IS) to evaluate policy updates from historical trajectories. However, IS-based RL inherently suffers from an exploration-stability dilemma: pure IS is vulnerable to gradient explosion and catastrophic training instability, while clipping severely limits the model's exploration capacity. By formalizing the concept of **Probability Capacity (Cap)**, we uncover a fundamental structural bottleneck: clipping constrains the update budget of high-reward but low-confidence reasoning paths, prematurely truncating gradients and stifling exploration. To address this challenge, we propose **Unbounded Positive Asymmetric Optimization (UPAO)**. UPAO applies an asymmetric scheme: it removes clipping and uses a self-anchored ratio via the stop-gradient operator for tokens with positive advantages, while retaining clipping for tokens with non-positive advantages to maintain stability. Experiments against 13 RL baselines show that UPAO improves reasoning performance across model families (Qwen3 and Olmo 3), architectures (Dense and MoE), and modalities (language and multimodal) without compromising training stability. These advantages persist as policy staleness increases.

1 INTRODUCTION

With the rapid advancement of large language models (LLMs), their ability to solve complex, multi-step mathematical and logical reasoning tasks has become increasingly prominent (Guo et al., 2025; Yang et al., 2025; Chen et al., 2025; BD Seed, 2026). However, navigating the vast and intricate search space of reasoning trajectories using supervised fine-tuning alone is often infeasible (Zelikman et al., 2022; Lightman et al., 2023). To address this challenge, reinforcement learning (RL) has emerged as the standard paradigm (Ouyang et al., 2022; Bai et al., 2022). While foundational on-policy methods like REINFORCE (Williams, 1992) offer stable optimization, they suffer from severe sample inefficiency. Consequently, modern RL frameworks such as Group Relative Policy Optimization (GRPO) (Shao et al., 2024) and Dynamic sAmpling Policy Optimization (DAPO) (Yu et al., 2025) have transitioned to multi-step optimization, which improves sample efficiency by utilizing *importance sampling (IS)* to evaluate policy updates from historical trajectories generated by π_{old}. More recently, a growing body of work has extended this paradigm along diverse directions, including sequence-level optimization via GSPO (Zheng et al., 2025a), critic-free global normalization via REINFORCE++ (Hu et al., 2025), and asymmetric importance sampling designs such as TOPR (Roux et al., 2025) and ASPO (Wang et al., 2025).

Despite the widespread adoption of IS-based RL algorithms, recent theoretical and empirical observations have identified a critical issue: *IS-based RL inherently suffers from the exploration-stability dilemma* (Liu et al., 2025; Zheng et al., 2025b; Yue et al., 2025; Ma et al., 2025). Specifically, pure unclipped IS is highly susceptible to pathological gradient explosion. When evaluating rare, long-tail reasoning paths, the IS ratio can explode, leading to catastrophic *training instability* (Schulman et al., 2017; 2015). To mitigate this instability, standard algorithms heavily rely on a *clip* mechanism, which forcibly bounds the policy update within a predefined trust region to preserve original model representations (Schulman et al., 2017; Li et al., 2026).

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

Although clipping safeguards optimization stability, it also limits the model's exploration capacity. While recent studies have recognized this limitation (Su et al., 2025; Gao et al., 2025; Zhang et al., 2025), a principled resolution remains elusive. To quantify this constraint, we define **Probability Capacity (Cap)** as the maximum allowable increase in π_θ for tokens with positive advantages, or the maximum allowable decrease for tokens with non-positive advantages, before the gradient is truncated. DAPO attempts to mitigate exploration stagnation with a larger upper clip bound ϵ_{high}. However, as shown in **Sec. 3**, clipping still bounds the Cap of high-reward, low-confidence reasoning tokens linearly in the behavior policy π_{old}. Even a small probability increase can push the importance sampling ratio beyond the threshold and truncate the gradient. Thus, adjusting ϵ_{high} alleviates but does not remove the structural bottleneck (Yue et al., 2025). Motivated by this observation, we ask:

***(Q)** How can we maximize a model's exploration while preventing training instability in RL?*

To address *(Q)*, we propose **Unbounded Positive Asymmetric Optimization (UPAO)**. UPAO uses an asymmetric scheme based on advantage signs (Zhu et al., 2025; Wang et al., 2025; Roux et al., 2025). For tokens with positive advantages, it removes clipping and uses a self-anchored ratio via the *stop-gradient operator (sg)* (Zhang et al., 2025). For tokens with non-positive advantages, it retains clipping to maintain stability. This design reinforces promising reasoning paths without sacrificing empirical stability. We summarize our **contributions** below:

• We define **Probability Capacity (Cap)**, uncovering a fundamental structural bottleneck in current RL algorithms: standard clipping inherently scales update budgets by π_{old}, prematurely truncating gradients and stifling critical exploration for rare, high-reward tokens.

• To resolve this limitation, we introduce **Unbounded Positive Asymmetric Optimization (UPAO)**, an algorithm that anchors the policy using a stop-gradient operator and employs an asymmetric design to substantially boost exploration capacity while guaranteeing rigorous training stability.

• Extensive experiments against 13 RL baselines demonstrate that UPAO consistently achieves superior accuracy across diverse model families (Qwen3, Olmo 3), network architectures (Dense, MoE), modalities, while robustly preserving convergence stability under severe policy staleness.

2 Related Work

Reinforcement Learning for LLM Reasoning. Building on REINFORCE (Williams, 1992), PPO (Schulman et al., 2017) became the standard RLHF/RLAIF paradigm (Ouyang et al., 2022; Bai et al., 2022), but its auxiliary critic limits scalability to long-context reasoning. GRPO (Shao et al., 2024) removes the critic via group-relative scoring and has powered breakthroughs in DeepSeek-R1 (Guo et al., 2025). DAPO (Yu et al., 2025) further introduces decoupled clip bounds, while GSPO (Zheng et al., 2025a) and related work (Mao et al., 2025; Liu et al., 2026) switch to sequence-level importance sampling (IS) for variance reduction. All these methods, however, still rely on restrictive clipping (Liu et al., 2025), which we argue structurally stifles exploration.

Importance Sampling and Asymmetric Optimization. TRPO (Schulman et al., 2015) and PPO introduced clipping to tame exploding IS ratios, and recent variants (BandPO (Li et al., 2026), SAPO (Gao et al., 2025), GMPO (Zhao et al., 2025), M2PO (Zheng et al., 2025b)) refine these bounds while still anchoring the IS ratio to the behavior policy π_{old}. This anchoring bottlenecks exploration in reasoning tasks where high-reward trajectories are sparse (Zelikman et al., 2022; Snell et al., 2024; Fan et al., 2026): gradients on rare high-reward paths are clipped before they can be reinforced. Asymmetric variants (e.g., ASPO (Wang et al., 2025), W-REINFORCE (Zhu et al., 2025), TOPR (Roux et al., 2025)) partially address this conflict by treating positive and non-positive advantages differently. UPAO advances this line by replacing π_{old} with a stop-gradient self-anchor $\text{sg}(\pi_\theta)$ for tokens with positive advantages while retaining clipping for tokens with non-positive advantages.

3 The Exploration-Stability Dilemma in Importance Sampling and Clipping

In this section, we analyze the mathematical framework of RL paradigms like REINFORCE and GRPO, exposing the structural dilemma between importance sampling-induced training instability and clipping-induced exploration stagnation in LLM reasoning.

3.1 Aggressive Updates Dilemma 1: Importance Sampling Induces Training Instability

The evolution of RL algorithms is fundamentally driven by the pursuit of sample efficiency and optimization stability. As a foundational approach, REINFORCE optimizes the policy strictly on-policy (Williams, 1992). Given a query q, its objective is the expected return $\mathcal{J}_{\text{REINFORCE}}(\theta) = \mathbb{E}_{q,\, o\sim\pi_\theta}[\mathcal{R}(q,o)]$, where $\mathcal{R}$ denotes the reward. Differentiating this objective and applying the log-derivative trick $\nabla_\theta \pi_\theta = \pi_\theta \nabla_\theta \log \pi_\theta$ to the sampling distribution yields the score-function form

$$\nabla_\theta \mathcal{J}_{\text{REINFORCE}}(\theta) = \mathbb{E}_{q,\, o\sim\pi_\theta}\Big[\hat{A}_{i,t}\, \nabla_\theta \log \pi_\theta(o_{i,t}|q, o_{i,<t})\Big], \tag{1}$$

where $\hat{A}_{i,t}$ is the advantage estimate for token (i,t) obtained from $\mathcal{R}$. For readability, **Eq. 1** and all subsequent gradient expressions in this section are written for a single token.

REINFORCE discards trajectories after a single update. To reuse a frozen rollout batch generated by the behavior policy π_{old}, the per-token importance sampling (IS) ratio $r_{i,t}(\theta) = \pi_\theta(o_{i,t}|q,o_{i,<t})/\pi_{\text{old}}(o_{i,t}|q,o_{i,<t})$ is used, where π_θ is the current policy. The per-token surrogate and its gradient are:

$$\mathcal{J}_{\text{IS}}(\theta) = \mathbb{E}_{q,o\sim\pi_{\text{old}}}\Big[\hat{A}_{i,t}\, r_{i,t}(\theta)\Big], \quad \nabla_\theta \mathcal{J}_{\text{IS}}(\theta) = \mathbb{E}_{q,o\sim\pi_{\text{old}}}\Big[\hat{A}_{i,t}\, r_{i,t}(\theta)\, \nabla_\theta \log \pi_\theta(o_{i,t}|q,o_{i,<t})\Big]. \tag{2}$$

Compared with the REINFORCE gradient in Eq. 1, the IS gradient in **Eq. 2** is scaled by $r_{i,t}(\theta)$. This ratio corrects for each token being sampled from π_{old}; holding prefixes and advantages fixed, the gradient approximates the on-policy one while $\pi_\theta \approx \pi_{\text{old}}$. The price is variance: $r_{i,t}(\theta) = \pi_\theta/\pi_{\text{old}}$ is unbounded above as $\pi_{\text{old}} \to 0$, so the importance weights are heavy-tailed. The problem worsens when the same rollout batch is reused for multiple updates, since π_θ drifts further from π_{old} with each update (policy staleness; Sec. 5.2). Consequently, a few low-probability tokens can dominate the gradient, causing gradient-norm spikes and **training instability**. Removing DAPO's upper clip ($\epsilon_{\text{high}} = \infty$) exposes this failure: its gradient norm explodes by many orders of magnitude (**Fig. 6(a)**).

3.2 Conservative Constraints Dilemma 2: Clipping Stifles Exploration

To prevent gradient explosion, standard algorithms such as GRPO (Shao et al., 2024) and DAPO (Yu et al., 2025) rely on clipping. We use DAPO with dual clip (Ye et al., 2020), as implemented in verl (Sheng et al., 2025). For G rollouts per query $q \sim \mathcal{Q}$, its objective is

$$\begin{aligned}\mathcal{J}_{\text{DAPO}}(\theta) = \mathbb{E}_{q\sim\mathcal{Q},\, \{o_i\}_{i=1}^{G}\sim\pi_{\text{old}}}\Big[&\frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|} \\ &\max\Big(\min\big(r_{i,t}\hat{A}_{i,t},\, \text{clip}(r_{i,t}, 1-\epsilon_{\text{low}}, 1+\epsilon_{\text{high}})\hat{A}_{i,t}\big),\, c\min(\hat{A}_{i,t}, 0)\Big)\Big],\end{aligned} \tag{3}$$

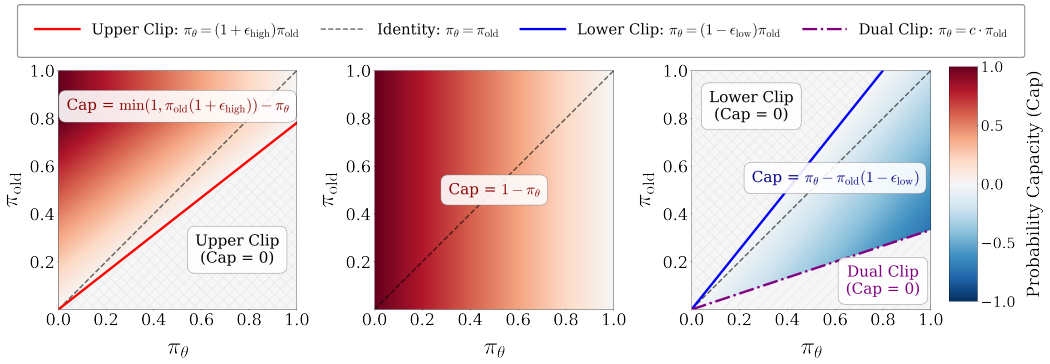
where $r_{i,t}$ abbreviates $r_{i,t}(\theta)$, ϵ_{low} and ϵ_{high} are clip bounds, and $c > 1$ is the dual-clip constant. The dual clip is active only for $\hat{A}_{i,t} < 0$, where $c\min(\hat{A}_{i,t}, 0) = c\hat{A}_{i,t}$ lower-bounds the surrogate; for $\hat{A}_{i,t} \geq 0$, the second argument is 0 and never exceeds the first. The core of **Eq. 3** is thus the clipped IS surrogate: for $\hat{A}_{i,t} > 0$, it reduces to $\mathcal{J}^{+}_{\text{DAPO}}(\theta) = \min[r_{i,t}(\theta), 1+\epsilon_{\text{high}}]\hat{A}_{i,t}$, with gradient

$$\nabla_\theta \mathcal{J}^{+}_{\text{DAPO}}(\theta) = \begin{cases}\hat{A}_{i,t} r_{i,t}(\theta) \nabla_\theta \log \pi_\theta(o_{i,t}|q,o_{i,<t}) & \text{if } r_{i,t}(\theta) \leq 1+\epsilon_{\text{high}} \\ 0 & \text{if } r_{i,t}(\theta) > 1+\epsilon_{\text{high}}\end{cases} \tag{4}$$

To analyze this gradient truncation (**Eq. 4**), we define **Probability Capacity (Cap)** as the maximum allowable increase (when $\hat{A} > 0$) or decrease (when $\hat{A} \leq 0$) in π_θ before the update is blocked. This capacity represents the effective update budget relative to the behavior policy π_{old}.

Fig. 1(a) shows that the gradient for tokens with positive advantages ($\hat{A}_{i,t} > 0$) vanishes when $r_{i,t}(\theta) > 1+\epsilon_{\text{high}}$. This constraint gives $\pi_\theta \leq (1+\epsilon_{\text{high}})\pi_{\text{old}}$. Because probabilities cannot exceed 1, the maximum achievable probability is $\min\big(1, (1+\epsilon_{\text{high}})\pi_{\text{old}}\big)$. The corresponding Cap is:

$$\text{Cap}(\pi_\theta, \pi_{\text{old}}) = \begin{cases}\min\big(1, (1+\epsilon_{\text{high}})\pi_{\text{old}}\big) - \pi_\theta & \text{if } \pi_\theta < (1+\epsilon_{\text{high}})\pi_{\text{old}} \\ 0 & \text{if } \pi_\theta \geq (1+\epsilon_{\text{high}})\pi_{\text{old}}\end{cases} \tag{5}$$



(a) DAPO ($\hat{A} > 0$) (b) UPAO ($\hat{A} > 0$) (c) DAPO/UPAO ($\hat{A} \leq 0$)

Figure 1: Visualization of the probability capacity (Cap) for DAPO and UPAO. Cap is defined as the maximum allowable increase in π_θ for tokens with positive advantages ($\hat{A} > 0$), or the maximum allowable decrease for tokens with non-positive advantages ($\hat{A} \leq 0$). (a) DAPO under $\hat{A} > 0$, where ϵ_{high} constrains the update, resulting in Cap $= 0$ within the Upper Clip region. (b) UPAO under $\hat{A} > 0$, where the absence of clipping allows all tokens to maintain a corresponding Cap. (c) DAPO and UPAO under $\hat{A} \leq 0$, where ϵ_{low} and c trigger Lower Clip and Dual Clip respectively, leading to Cap $= 0$ in those regions.

This piecewise formulation reveals a severe structural bottleneck: the update budget is linearly dependent on π_{old}. For low-likelihood reasoning tokens (e.g., $\pi_{\text{old}} = 0.01$), a standard threshold of $\epsilon_{\text{high}} = 0.28$ restricts the maximum absolute probability increase to a mere 0.0028. Once π_θ reaches 0.0128, the capacity drops to exactly zero (entering the "Upper Clip" region shown in Fig. 1(a)), and the gradient is nullified. Even if this specific action yields an exceptionally high advantage, its effective update is prematurely blocked. Ultimately, this mechanism structurally **stifles exploration**, preventing the algorithm from reinforcing promising long-tail trajectories and leaving optimal reasoning paths undiscovered.

Key Insights of Sec. 3: The Exploration-Stability Dilemma

(1) IS improves sample efficiency but destabilizes training. REINFORCE is strictly on-policy and discards each rollout batch after a single update; IS instead reuses a batch from π_{old} by scaling each token's gradient by $r_{i,t}(\theta)$ (Eq. 2). Since $r_{i,t}(\theta)$ is unbounded above, tokens with small π_{old} can dominate the update and destabilize training.

(2) Clipping restores stability but limits exploration. Clipping zeroes the gradient of a token with positive advantage once $\pi_\theta > (1+\epsilon_{\text{high}})\pi_{\text{old}}$ (Eq. 4), so its Cap, $\min\left(1, (1+\epsilon_{\text{high}})\pi_{\text{old}}\right) - \pi_\theta$, scales with π_{old} (Fig. 1(a)). Rare tokens thus get little room to be reinforced.

(3) The resulting dilemma. Without clipping, IS is unstable; with clipping, it stifles exploration. Resolving this trade-off is the goal of Sec. 4.

4 Breaking the Dilemma: Unbounded Positive Asymmetric Optimization

We resolve the dilemma with two components, which together form **Unbounded Positive Asymmetric Optimization (UPAO)**: a self-anchored ratio that replaces π_{old} with $\text{sg}(\pi_\theta)$ (Sec. 4.1), and an asymmetric objective that retains clipping only for tokens with non-positive advantages (Sec. 4.2).

4.1 Unbounded Formulation for Positive Advantages

To eliminate the pathological gradients introduced by π_{old} while retaining multi-step optimization over sampled data, for tokens with positive advantages ($\hat{A}_{i,t} > 0$), we replace the standard importance

sampling ratio $r_{i,t}(\theta)$ with a self-anchored ratio $\tilde{r}_{i,t}(\theta)$. Using the **stop-gradient operator (sg)**, we define the unbounded positive component $\mathcal{J}^{+}_{\text{UPAO}}(\theta)$:

$$\mathcal{J}^{+}_{\text{UPAO}}(\theta) = \mathbb{E}_{q\sim\mathcal{Q},\,\{o_i\}_{i=1}^{G}\sim\pi_{\text{old}}}\left[\frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}\hat{A}_{i,t}\,\tilde{r}_{i,t}(\theta)\right], \quad \hat{A}_{i,t}>0 \tag{6}$$

where $\tilde{r}_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\text{sg}(\pi_\theta(o_{i,t}|q,o_{i,<t}))}$. As $\text{sg}(\pi_\theta)$ is constant in backpropagation and equals π_θ in value, the log-derivative trick gives

$$\begin{aligned}\nabla_\theta\mathcal{J}^{+}_{\text{UPAO}}(\theta) &= \mathbb{E}_{q,\{o_i\}\sim\pi_{\text{old}}}\left[\frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i,t}\hat{A}_{i,t}\frac{\nabla_\theta\pi_\theta(o_{i,t}|q,o_{i,<t})}{\text{sg}(\pi_\theta(o_{i,t}|q,o_{i,<t}))}\right]\\ &= \mathbb{E}_{q,\{o_i\}\sim\pi_{\text{old}}}\left[\frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i,t}\hat{A}_{i,t}\,\nabla_\theta\log\pi_\theta(o_{i,t}|q,o_{i,<t})\right]\end{aligned} \tag{7}$$

Eq. 7 anchors the policy to π_θ rather than π_{old}, removing the root cause of IS-induced instability. As shown in **Fig. 1(b)**, the Cap is no longer tied to π_{old}: every token keeps the full budget Cap $= 1-\pi_\theta$.

4.2 Asymmetric Optimization

Building on this formulation, UPAO addresses the dilemma analyzed in the previous section with an asymmetric objective: tokens with positive advantages use the unbounded component $\mathcal{J}^{+}_{\text{UPAO}}$ to promote exploration and reinforce rare, low-confidence tokens, whereas tokens with non-positive advantages retain clipping as a safeguard against overly aggressive penalization (**Fig. 1(c)**; see **Appendix A**). The full objective of UPAO is:

$$\begin{aligned}\mathcal{J}_{\text{UPAO}}(\theta) = \mathbb{E}_{q\sim\mathcal{Q},\,\{o_i\}_{i=1}^{G}\sim\pi_{\text{old}}}\Bigg[&\frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}\\ &\begin{cases}\hat{A}_{i,t}\,\tilde{r}_{i,t}(\theta) & \text{if } \hat{A}_{i,t}>0\\ \max\Big(\min\big(r_{i,t}\hat{A}_{i,t},\,\text{clip}(r_{i,t},1-\epsilon_{\text{low}},1+\epsilon_{\text{high}})\hat{A}_{i,t}\big),\,c\hat{A}_{i,t}\Big) & \text{if } \hat{A}_{i,t}\le 0\end{cases}\Bigg],\end{aligned} \tag{8}$$

where the non-positive branch, including $r_{i,t}$ and $c=3$, coincides with DAPO's objective (Eq. 3) restricted to $\hat{A}_{i,t}\le 0$; for a shared group-relative rollout advantage, $\hat{A}_{i,t}=\hat{A}_i$ for all t.

Key Insights of Sec. 4: How UPAO Breaks the Dilemma

(1) Self-anchoring frees positive tokens. For tokens with positive advantages, the self-anchored ratio $\tilde{r}_{i,t}(\theta)$ removes π_{old} from the gradient weight, avoiding instability from small π_{old}, and raises the Cap to $1-\pi_\theta$, so rare tokens retain ample room to be reinforced.

(2) Asymmetry resolves the dilemma. Tokens with non-positive advantages retain DAPO's clipping and dual clip, keeping penalization conservative and training stable.

5 Experiments

5.1 Experimental Setup

Models. We evaluate UPAO across diverse model families, including the Qwen3 suite (Yang et al., 2025) spanning dense LLMs in base (Qwen3-14B-Base) and instruct (Qwen3-8B) variants, an MoE model (Qwen3-30B-A3B-Base), and a vision-language model (Qwen3-VL-8B-Instruct), as well as Olmo-3-7B-Instruct-SFT to verify cross-family generalization.

Training and Evaluation. We adopt three protocols: (i) following Liu et al. (2025), we train Qwen3-8B and Olmo-3-7B-Instruct-SFT on MATH (Levels 3–5) (Hendrycks et al., 2021) and evaluate on

AIME24 (Mathematical Association of America, 2024), AMC23, MATH500, Minerva (Lewkowycz et al., 2022), OlympiadBench (He et al., 2024), MMLU-Pro (Wang et al., 2024), and BBH (Suzgun et al., 2023); (ii) following Yu et al. (2025), we train Qwen3-14B-Base and Qwen3-30B-A3B-Base on DAPO-17K-MATH and evaluate on AIME24; and (iii) following Zhao et al. (2025), we train Qwen3-VL-8B-Instruct on the Geometry3K (Lu et al., 2021) training set and evaluate on its test set. Beyond accuracy, we assess exploration using Pass@K and generation entropy, and stability using KL divergence and gradient norms. Higher Pass@K and entropy indicate better solution discovery and policy diversity, while smoother KL and gradient-norm curves indicate more stable training.

Algorithms and Baselines. We compare against thirteen RL baselines: DAPO (Yu et al., 2025), GRPO (Shao et al., 2024), GMPO (Zhao et al., 2025), ASPO (Wang et al., 2025), CISPO (Chen et al., 2025), Dr. GRPO (Liu et al., 2025), W-REINFORCE (Zhu et al., 2025), REINFORCE++ (Hu et al., 2025), RLOO (Ahmadian et al., 2024), DPPO (Qi et al., 2026), GSPO (Zheng et al., 2025a), SAPO (Gao et al., 2025), and TOPR (Roux et al., 2025).

Implementation Details. Our implementation is based on the verl framework (Sheng et al., 2025), utilizing vLLM for efficient rollout generation and evaluation (Kwon et al., 2023). Comprehensive details are provided in **Appendix B**.

5.2 Main Results

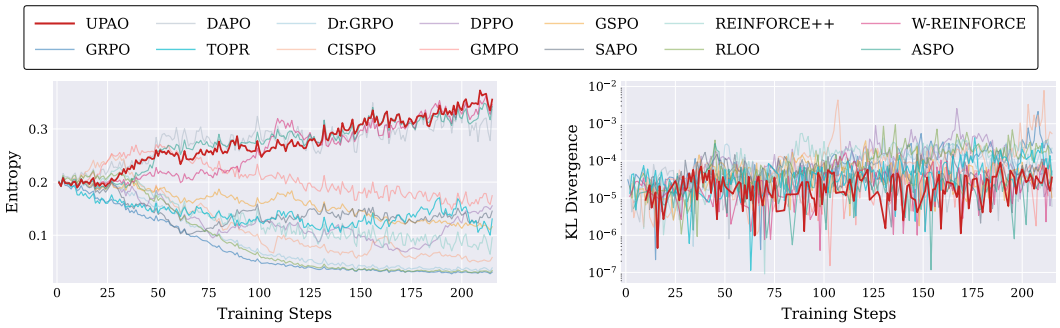
Method	In-Domain					Out-of-Domain		Average	Rank
	AIME24	AMC23	MATH500	Minerva	Olympiad	MMLU-Pro	BBH		
GRPO (Shao et al., 2024)	35.73	75.00	86.00	30.88	51.34	63.08	73.00	59.29	12
DAPO (Yu et al., 2025)	39.06	85.00	**88.20**	29.78	55.80	65.31	77.71	62.98	6
DPPO (Qi et al., 2026)	40.10	87.50	86.20	30.51	53.27	65.66	77.13	62.91	7
Dr. GRPO (Liu et al., 2025)	33.33	85.00	85.80	30.15	51.19	65.41	74.81	60.81	10
CISPO (Chen et al., 2025)	38.02	87.50	86.60	29.04	55.65	66.86	76.21	62.84	8
GMPO (Zhao et al., 2025)	37.50	87.50	87.00	31.25	55.06	66.19	78.11	63.23	4
GSPO (Zheng et al., 2025a)	**40.52**	85.00	**88.20**	31.25	55.80	65.09	77.48	63.33	3
SAPO (Gao et al., 2025)	39.90	82.50	87.20	30.88	55.65	64.76	75.32	62.32	9
REINFORCE++ (Hu et al., 2025)	20.52	62.50	78.80	29.78	42.26	61.81	71.22	52.41	14
RLOO (Ahmadian et al., 2024)	31.67	80.00	85.20	28.68	50.60	63.46	73.61	59.03	13
W-REINFORCE (Zhu et al., 2025)	35.52	80.00	85.80	30.15	53.27	64.09	74.92	60.54	11
ASPO (Wang et al., 2025)	37.50	85.00	87.60	29.78	58.48	67.14	77.88	63.34	2
TOPR (Roux et al., 2025)	40.00	87.50	87.60	30.15	55.80	63.34	78.11	63.21	5
UPAO (Ours)	**40.52**	**90.00**	87.60	**33.09**	**59.08**	**69.04**	**79.13**	**65.49**	**1**

Table 1: Performance of UPAO against thirteen RL baselines on five in-domain and two out-of-domain benchmarks. All algorithms are trained on MATH (Levels 3–5) using Qwen3-8B following Liu et al. (2025). Performance is evaluated by Pass@1 accuracy (%), and **Average** is computed across all seven benchmarks. The best score in each column is in **bold**.

Overall Performance. We benchmark UPAO against thirteen representative RL baselines under the unified protocol of Liu et al. (2025), training on MATH (Levels 3–5) with Qwen3-8B and evaluating Pass@1 accuracy across five in-domain and two out-of-domain benchmarks. As reported in **Table 1**, UPAO ranks first with an average accuracy of 65.49%, outperforming the strongest baseline, ASPO, at 63.34%. UPAO also achieves the best results on AMC23, Minerva, OlympiadBench, MMLU-Pro, and BBH, and ties GSPO for the best result on AIME24. These gains across both in-domain and out-of-domain benchmarks demonstrate the broad effectiveness of the unbounded positive update.

Exploration Capacity and Training Stability. We further compare UPAO with thirteen baselines throughout training. **Fig. 2(a)** reveals a clear bifurcation in entropy. GRPO, Dr. GRPO, RLOO, CISPO, and REINFORCE++ collapse sharply, GSPO and GMPO decay steadily, and DPPO, SAPO, and TOPR settle below their initial entropy. W-REINFORCE, ASPO, and UPAO, which treat positive and non-positive advantages asymmetrically, instead increase entropy, and UPAO ends highest. Consistent with the Cap analysis, removing the π_{old}-dependent cutoff for tokens with positive advantages lets UPAO keep reinforcing rare tokens and sustain exploration. **Fig. 2(b)** further shows that UPAO's KL divergence remains among the lowest throughout training, comparable to ASPO, W-REINFORCE, and GMPO, and below GRPO, RLOO, CISPO, DPPO, and REINFORCE++ late in training. These results show that UPAO preserves exploration while maintaining stable optimization.

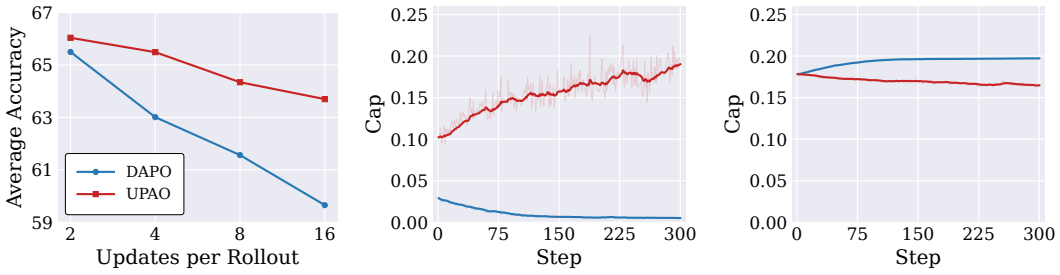
Robustness to Policy Staleness. We evaluate DAPO and UPAO across different levels of policy staleness, quantified by the number m of optimizer updates performed on each rollout batch: more



(a) Entropy (b) KL divergence

Figure 2: Exploration capacity and training stability of UPAO and thirteen RL baselines on Qwen3-8B during training: (a) entropy and (b) KL divergence between the active policy and the reference model on MATH (Levels 3–5).

updates cause the current policy π_θ to drift further away from the behavior policy π_{old} (**Table A5**). As shown in **Fig. 3(a)**, UPAO consistently outperforms DAPO and suffers a markedly smaller accuracy drop as m increases from 2 to 16: DAPO falls from 65.50% to 59.66% (-5.84), whereas UPAO falls only from 66.04% to 63.70% (-2.34). **Fig. 3(b)** and **Fig. 3(c)** explain this robustness through Cap. For $\hat{A} > 0$, removing clipping allows UPAO to maintain a substantially larger Cap than DAPO, releasing update budget to reinforce promising long-tail trajectories and enhance exploration. Conversely, for $\hat{A} \leq 0$, UPAO preserves DAPO's clipping to prevent destructive updates on reversed gradients and maintain training stability. **Appendix C** formalizes this: relative to the on-policy gradient, DAPO discards a token with $\hat{A} > 0$ once its ratio exceeds $1 + \epsilon_{\text{high}}$, whereas UPAO keeps it with weight $1/r$. Under a stylized model where each update raises such tokens' log-probability by a constant δ, DAPO discards more of them as m grows, so UPAO's retained signal overtakes DAPO's by a ratio rising toward a limit that tends to $1/\log(1 + \epsilon_{\text{high}}) \approx 4.05$ for $\epsilon_{\text{high}} = 0.28$ as $\delta \to 0$. This explains the direction, not the magnitude, of the widening gap in Fig. 3(a).

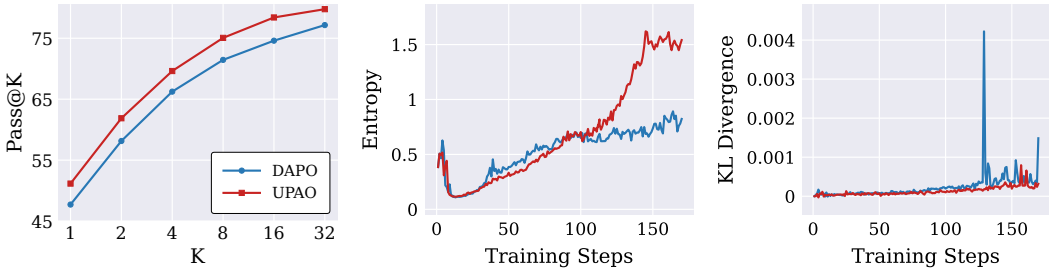


(a) Average Accuracy (b) Cap($\hat{A} > 0$) (c) Cap($\hat{A} \leq 0$)

Figure 3: Robustness of DAPO and UPAO to policy staleness. (a) Average accuracy across seven benchmarks at different numbers of updates per rollout; (b) Cap for tokens with positive advantages under 4 updates per rollout during training; and (c) Cap for tokens with non-positive advantages under 4 updates per rollout.

5.3 Generalization across Models and Modalities

Larger Models. To examine whether UPAO generalizes to larger models, we evaluate it on **Qwen3-14B-Base** under protocol (ii), with DAPO and UPAO sharing the configuration in **Table A1**. UPAO consistently outperforms DAPO across all values of K in **Fig. 4(a)**, from 47.71% to 51.15% at $K = 1$ and from 77.16% to 79.77% at $K = 32$, so the gain persists at large K. Its entropy (**Fig. 4(b)**) tracks DAPO's for $\sim$90 steps, then overtakes it and ends at about 1.5 versus 0.8, as unbounded positive updates keep reinforcing rewarded low-probability tokens. Its KL divergence stays below 10^{-3}, while DAPO briefly spikes to about 4×10^{-3} (**Fig. 4(c)**), and its gradient norm stays comparable to



(a) Pass@K (b) Entropy (c) KL divergence

Figure 4: Exploration capacity and stability of DAPO and UPAO on Qwen3-14B-Base trained on DAPO-17K-MATH: (a) Pass@K on AIME24, (b) entropy of generation probabilities on DAPO-17K-MATH, and (c) KL divergence between the policy and the reference model on DAPO-17K-MATH.

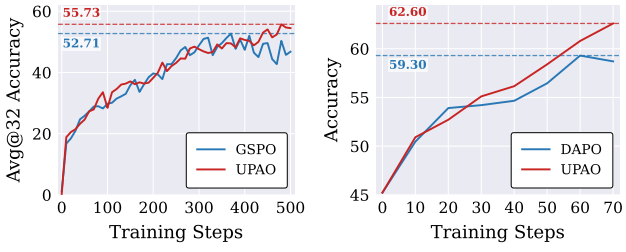
DAPO's (**Fig. A1(b)**). Thus, UPAO improves exploration and performance on a larger model without compromising training stability, since tokens with non-positive advantages remain clipped.

MoE Architectures. We next evaluate UPAO on the Qwen3-30B-A3B-Base MoE model, where dynamic expert routing introduces additional variance into policy updates. UPAO keeps GSPO's lower clip $\epsilon_{\text{low}} = 3 \times 10^{-4}$ for non-positive advantages and applies the self-anchored ratio to positive ones (**Table A2**). As shown in **Fig. 5(a)**, UPAO reaches 55.73% Avg@32 accuracy on AIME24, improving over GSPO by 3.02 points. The curves stay close for most of training; the gap opens late, once GSPO plateaus after peaking at 52.71% while UPAO keeps improving, suggesting that unbounded positive updates matter most when the clipped baseline saturates. Its KL divergence remains of the same order as GSPO's throughout training, with both below 5×10^{-4} (**Fig. A2(a)**), demonstrating that UPAO transfers to sparse MoE architectures without destabilizing optimization.

Model Families. To evaluate cross-family generalization, we compare UPAO with DAPO on Olmo-3-7B-Instruct-SFT trained on MATH (Levels 3–5) under protocol (i) (**Table A3**). UPAO improves over DAPO on all seven in-domain and out-of-domain benchmarks and raises average accuracy from 44.76% to 46.12% (**Table 2**). Gains are largest on AMC23 (+2.50) and MMLU-Pro (+2.24) and small on OlympiadBench (+0.30) and BBH (+0.42). This consistent improvement shows that the gains of UPAO are not specific to the Qwen family.

Vision-Language Models. Finally, we evaluate UPAO on Qwen3-VL-8B-Instruct trained on Geometry3K (Lu et al., 2021) to test whether it transfers beyond language-only reasoning (protocol (iii); **Table A4**). From the same initial accuracy, UPAO reaches 62.60% accuracy on the Geometry3K test set, improving over DAPO's best of 59.30% by 3.30 points (**Fig. 5(b)**); DAPO peaks at step 60, whereas UPAO is still improving at the last step. Both keep a similar KL divergence below 1.5×10^{-4} (**Fig. A2(b)**), so the gain comes without additional policy drift.

Together, these results demonstrate that UPAO generalizes across model scales, architectures, families, and modalities: in each setting it improves accuracy over the clipped baseline at a comparable KL divergence, matching the split between unbounded positive updates and clipping for the rest.



(a) Qwen3-30B-A3B-Base (b) Qwen3-VL-8B-Instruct

Figure 5: (a) UPAO vs. GSPO on Qwen3-30B-A3B-Base (AIME24 Avg@32); (b) UPAO vs. DAPO on Qwen3-VL-8B-Instruct (Geometry3K accuracy).

Domain	Benchmark	Method	
		DAPO	UPAO
In-Domain	AIME24	17.60	**19.17**
	AMC23	67.50	**70.00**
	MATH500	81.20	**82.60**
	Minerva	22.79	**23.90**
	Olympiad	48.21	**48.51**
Out-of-Domain	MMLU-Pro	34.46	**36.70**
	BBH	41.51	**41.93**
	Average	44.76	**46.12**

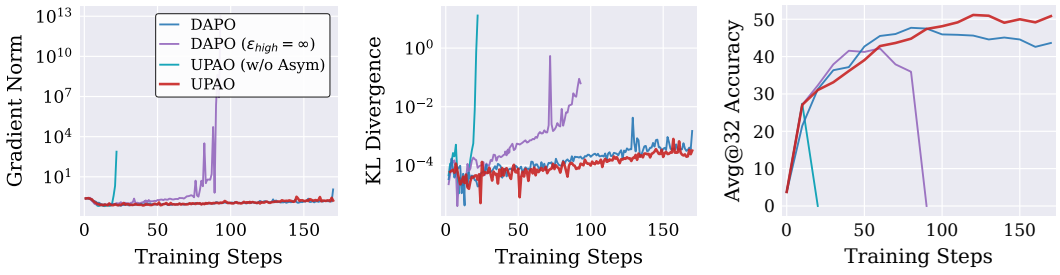
Table 2: UPAO vs. DAPO on Olmo-3-7B-Instruct-SFT trained on MATH (Levels 3–5), evaluated by accuracy.

5.4 Ablation Studies on Key Design

Necessity of the Self-Anchored Ratio. We first ask whether DAPO can recover UPAO's exploration by simply relaxing the trust region of its standard ratio $r_{i,t}(\theta)$, namely setting $\epsilon_{\text{high}} = \infty$, which turns its positive branch into the unclipped IS term $\frac{\pi_\theta}{\pi_{\text{old}}}\hat{A}$ (**Table 3**). As shown by the purple curve in **Fig. 6(b)**, this variant becomes unstable later in training as its KL divergence rises rapidly, ultimately leading to training failure and model collapse with accuracy plunging to 0% by step 90 (**Fig. 6(c)**). In contrast, UPAO maintains a gradient norm comparable to DAPO throughout training (Fig. 6(a)) while keeping its KL divergence low. This confirms that simply removing the upper clip while still anchoring the ratio to π_{old} is insufficient, and that the stop-gradient operator in our self-anchored ratio $\tilde{r}_{i,t}(\theta)$ (**Eq. 6**) is what makes unbounded positive updates safe.

	$\hat{A}_{i,t} > 0$	$\hat{A}_{i,t} \le 0$
DAPO	$\min\big(\frac{\pi_\theta}{\pi_{\text{old}}}, 1+\epsilon_{\text{high}}\big)\hat{A}$	$\max\big(\min\big(\frac{\pi_\theta}{\pi_{\text{old}}}\hat{A}, \text{clip}\big(\frac{\pi_\theta}{\pi_{\text{old}}}, 1-\epsilon_{\text{low}}, 1+\epsilon_{\text{high}}\big)\hat{A}\big), c\hat{A}\big)$
DAPO ($\epsilon_{\text{high}} = \infty$)	$\frac{\pi_\theta}{\pi_{\text{old}}}\hat{A}$	$\max\big(\min\big(\frac{\pi_\theta}{\pi_{\text{old}}}\hat{A}, \text{clip}\big(\frac{\pi_\theta}{\pi_{\text{old}}}, 1-\epsilon_{\text{low}}, 1+\epsilon_{\text{high}}\big)\hat{A}\big), c\hat{A}\big)$
UPAO (w/o Asym)	$\frac{\pi_\theta}{\text{sg}(\pi_\theta)}\hat{A}$	$\frac{\pi_\theta}{\text{sg}(\pi_\theta)}\hat{A}$
UPAO	$\frac{\pi_\theta}{\text{sg}(\pi_\theta)}\hat{A}$	$\max\big(\min\big(\frac{\pi_\theta}{\pi_{\text{old}}}\hat{A}, \text{clip}\big(\frac{\pi_\theta}{\pi_{\text{old}}}, 1-\epsilon_{\text{low}}, 1+\epsilon_{\text{high}}\big)\hat{A}\big), c\hat{A}\big)$

Table 3: Per-token objectives of the ablated methods, where π_θ and π_{old} denote $\pi_\theta(o_{i,t}|q, o_{i,<t})$ and $\pi_{\text{old}}(o_{i,t}|q, o_{i,<t})$, so $\pi_\theta/\pi_{\text{old}} = r_{i,t}(\theta)$ is the IS ratio; sg is the stop-gradient operator, $\hat{A} = \hat{A}_{i,t}$, and c is the dual-clip constant.



(a) Gradient norm (b) KL divergence (c) Accuracy

Figure 6: Training stability and ablations on Qwen3-14B-Base trained on DAPO-17K-MATH: (a) gradient norms and (b) KL divergence for DAPO, UPAO, DAPO with $\epsilon_{\text{high}} = \infty$, and UPAO (w/o Asym); and (c) Avg@32 accuracy on AIME24 for the same four methods.

Necessity of the Asymmetric Design. We further evaluate a symmetric variant, UPAO (w/o Asym), that applies unbounded updates to tokens with both positive and non-positive advantages, i.e., an off-policy REINFORCE-style objective without clipping (Table 3). As shown by the cyan curve in Fig. 6(b), this variant collapses within the first 25 steps as its KL divergence surges, resulting in training failure and a performance collapse with 0% accuracy by step 20 (Fig. 6(c)). Hence, while unbounded updates benefit tokens with positive advantages, DAPO-style clipping for tokens with non-positive advantages in **Eq. 8** remains an indispensable safeguard.

6 Conclusion

To address the exploration-stability dilemma in IS-based reinforcement learning, we introduce Unbounded Positive Asymmetric Optimization (UPAO). Probability Capacity (Cap) shows how the stop-gradient self-anchor bypasses conservative clipping constraints for tokens with positive advantages. Experiments against 13 RL baselines confirm that UPAO improves reasoning performance across model families (Qwen3 and Olmo 3), architectures (Dense and MoE), and modalities (language and multimodal) while maintaining training stability. We discuss the broader impact and limitations of UPAO in **Appendix E** and **Appendix F**, respectively.

AI USE STATEMENT

AI-assisted tools were used primarily to improve the clarity, grammar, and presentation of the manuscript. The authors remain fully responsible for the scientific content, technical claims, and final text.

REFERENCES

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting reinforce-style optimization for learning from human feedback in llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12248–12267, 2024.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

BD Seed. Seed 2.0 model card: Towards intelligence frontier for real-world complexity. Technical report, ByteDance, 2026. URL `https://seed.bytedance.com/en/seed2`.

Aili Chen, Aonian Li, Bangwei Gong, Binyang Jiang, Bo Fei, Bo Yang, Boji Shan, Changqing Yu, Chao Wang, Cheng Zhu, et al. Minimax-m1: Scaling test-time compute efficiently with lightning attention. *arXiv preprint arXiv:2506.13585*, 2025.

Chongyu Fan, Yihua Zhang, Jinghan Jia, Alfred O. Hero, and Sijia Liu. Cyclicreflex: Improving reasoning models via cyclical reflection token scheduling. In *The Fourteenth International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=4oOF4J2xSy`.

Chang Gao, Chujie Zheng, Xiong-Hui Chen, Kai Dang, Shixuan Liu, Bowen Yu, An Yang, Shuai Bai, Jingren Zhou, and Junyang Lin. Soft adaptive policy optimization. *arXiv preprint arXiv:2511.20347*, 2025.

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

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. 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)*, pp. 3828–3850, 2024.

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.

Jian Hu, Jason Klein Liu, Haotian Xu, and Wei Shen. Reinforce++: Stabilizing critic-free policy optimization with global advantage normalization. *arXiv preprint arXiv:2501.03262*, 2025.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pp. 611–626, 2023.

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

Yuan Li, Bo Wang, Yufei Gao, Yuqian Yao, Xinyuan Wang, Zhangyue Yin, and Xipeng Qiu. Bandpo: Bridging trust regions and ratio clipping via probability-aware bounds for llm reinforcement learning. *arXiv preprint arXiv:2603.04918*, 2026.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The twelfth international conference on learning representations*, 2023.

Fanfan Liu, Youyang Yin, Peng Shi, Siqi Yang, Zhixiong Zeng, and Haibo Qiu. Length-unbiased sequence policy optimization: Revealing and controlling response length variation in rlvr. *arXiv preprint arXiv:2602.05261*, 2026.

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.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-GPS: Interpretable geometry problem solving with formal language and symbolic reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pp. 6774–6786, 2021.

Wenhan Ma, Hailin Zhang, Liang Zhao, Yifan Song, Yudong Wang, Zhifang Sui, and Fuli Luo. Stabilizing moe reinforcement learning by aligning training and inference routers. *arXiv preprint arXiv:2510.11370*, 2025.

Hanyi Mao, Quanjia Xiao, Lei Pang, and Haixiao Liu. Clip your sequences fairly: Enforcing length fairness for sequence-level rl. *arXiv preprint arXiv:2509.09177*, 2025.

Mathematical Association of America. American Invitational Mathematics Examination (AIME) 2024, 2024.

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.

Penghui Qi, Xiangxin Zhou, Zichen Liu, Tianyu Pang, Chao Du, Min Lin, and Wee Sun Lee. Rethinking the trust region in llm reinforcement learning. *arXiv preprint arXiv:2602.04879*, 2026.

Nicolas Le Roux, Marc G Bellemare, Jonathan Lebensold, Arnaud Bergeron, Joshua Greaves, Alex Fréchette, Carolyne Pelletier, Eric Thibodeau-Laufer, Sándor Toth, and Sam Work. Tapered off-policy reinforce: Stable and efficient reinforcement learning for llms. *arXiv preprint arXiv:2503.14286*, 2025.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897. PMLR, 2015.

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

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

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

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

Zhenpeng Su, Leiyu Pan, Xue Bai, Dening Liu, Guanting Dong, Jiaming Huang, Wenping Hu, Fuzheng Zhang, Kun Gai, and Guorui Zhou. Klear-reasoner: Advancing reasoning capability via gradient-preserving clipping policy optimization. *arXiv preprint arXiv:2508.07629*, 2025.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed H Chi, Denny Zhou, et al. Challenging big-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13003–13051, 2023.

Jiakang Wang, Runze Liu, Lei Lin, Wenping Hu, Xiu Li, Fuzheng Zhang, Guorui Zhou, and Kun Gai. Aspo: Asymmetric importance sampling policy optimization. *arXiv preprint arXiv:2510.06062*, 2025.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37: 95266–95290, 2024.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

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

Deheng Ye, Zhao Liu, Mingfei Sun, Bei Shi, Peilin Zhao, Hao Wu, Hongsheng Yu, Shaojie Yang, Xipeng Wu, Qingwei Guo, et al. Mastering complex control in moba games with deep reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 6672–6679, 2020.

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.

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.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488, 2022.

Yifan Zhang, Yifeng Liu, Huizhuo Yuan, Yang Yuan, Quanquan Gu, and Andrew Chi-Chih Yao. On the design of kl-regularized policy gradient algorithms for llm reasoning. *arXiv preprint arXiv:2505.17508*, 2025.

Yuzhong Zhao, Yue Liu, Junpeng Liu, Jingye Chen, Xun Wu, Yaru Hao, Tengchao Lv, Shaohan Huang, Lei Cui, Qixiang Ye, et al. Geometric-mean policy optimization. *arXiv preprint arXiv:2507.20673*, 2025.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025a.

Haizhong Zheng, Jiawei Zhao, and Beidi Chen. Prosperity before collapse: How far can off-policy rl reach with stale data on llms? *arXiv preprint arXiv:2510.01161*, 2025b.

Xinyu Zhu, Mengzhou Xia, Zhepei Wei, Wei-Lin Chen, Danqi Chen, and Yu Meng. The surprising effectiveness of negative reinforcement in llm reasoning. *arXiv preprint arXiv:2506.01347*, 2025.

APPENDIX

APPENDIX CONTENTS

A Derivation of Probability Capacity for Negative Advantages **14**

B Training and Evaluation Details **15**

C Why UPAO Retains More Gradient Signal for Tokens with Positive Advantages under High Policy Staleness **17**

C.1 Setup 17

C.2 Retained Signal under Policy Drift 17

D Exploration and Stability Curves **19**

E Broader Impact **20**

F Limitations **20**

A Derivation of Probability Capacity for Negative Advantages

The non-positive branch of Eq. 8 contains the complete clipped and dual-clipped surrogate for both UPAO and DAPO; it coincides with Eq. 3 restricted to $\hat{A}_{i,t} \le 0$, where $c\min(\hat{A}_{i,t}, 0) = c\hat{A}_{i,t}$. As in that equation, write $r_{i,t} = r_{i,t}(\theta)$ below and consider a token with $\hat{A}_{i,t} < 0$; the case $\hat{A}_{i,t} = 0$ is trivial because both the surrogate and its gradient are zero.

First consider the inner minimum. Multiplication by $\hat{A}_{i,t} < 0$ reverses order, so

$$\begin{aligned} &\min\left(r_{i,t}\hat{A}_{i,t}, \text{clip}(r_{i,t}, 1-\epsilon_{\text{low}}, 1+\epsilon_{\text{high}})\hat{A}_{i,t}\right) \\ &= \hat{A}_{i,t}\max\left(r_{i,t}, \text{clip}(r_{i,t}, 1-\epsilon_{\text{low}}, 1+\epsilon_{\text{high}})\right) \\ &= \begin{cases} (1-\epsilon_{\text{low}})\hat{A}_{i,t} & \text{if } r_{i,t} < 1-\epsilon_{\text{low}}, \\ r_{i,t}\hat{A}_{i,t} & \text{if } r_{i,t} \ge 1-\epsilon_{\text{low}}. \end{cases} \end{aligned} \tag{A1}$$

For $r_{i,t} > 1+\epsilon_{\text{high}}$, the maximum in **Eq. A1** selects $r_{i,t}$, so the upper clip is inactive for negative advantages.

Next apply the outer maximum with $c\hat{A}_{i,t}$. Since $\max(r_{i,t}\hat{A}_{i,t}, c\hat{A}_{i,t}) = \hat{A}_{i,t}\min(r_{i,t}, c)$ and $c > 1 > 1-\epsilon_{\text{low}}$, the complete per-token surrogate is

$$\begin{aligned} &\max\left(\min\left(r_{i,t}\hat{A}_{i,t}, \text{clip}(r_{i,t}, 1-\epsilon_{\text{low}}, 1+\epsilon_{\text{high}})\hat{A}_{i,t}\right), c\hat{A}_{i,t}\right) \\ &= \begin{cases} (1-\epsilon_{\text{low}})\hat{A}_{i,t} & \text{if } r_{i,t} < 1-\epsilon_{\text{low}}, \\ r_{i,t}\hat{A}_{i,t} & \text{if } 1-\epsilon_{\text{low}} \le r_{i,t} \le c, \\ c\hat{A}_{i,t} & \text{if } r_{i,t} > c. \end{cases} \end{aligned} \tag{A2}$$

Differentiating **Eq. A2** gives

$$\nabla_\theta(\text{negative-branch surrogate}) = \begin{cases} 0 & \text{if } r_{i,t} < 1-\epsilon_{\text{low}}, \\ \hat{A}_{i,t} r_{i,t} \nabla_\theta \log \pi_\theta(o_{i,t}|q, o_{i,<t}) & \text{if } 1-\epsilon_{\text{low}} \le r_{i,t} \le c, \\ 0 & \text{if } r_{i,t} > c. \end{cases} \tag{A3}$$

Thus **Eq. A3** is non-zero only in the effective window

$$(1-\epsilon_{\text{low}}) \le r_{i,t} \le c. \tag{A4}$$

Writing π_θ and π_{old} for the current and behavior probabilities of token (i,t), respectively, multiplying **Eq. A4** by π_{old} gives

$$(1-\epsilon_{\text{low}})\pi_{\text{old}} \le \pi_\theta \le c\pi_{\text{old}}. \tag{A5}$$

Because an update with $\hat{A}_{i,t} < 0$ decreases π_θ, the lower boundary in **Eq. A5** leaves the remaining update room $\pi_\theta - (1-\epsilon_{\text{low}})\pi_{\text{old}}$. Therefore,

$$\text{Cap}(\pi_\theta, \pi_{\text{old}}) = \begin{cases} \pi_\theta - (1-\epsilon_{\text{low}})\pi_{\text{old}} & \text{if } (1-\epsilon_{\text{low}})\pi_{\text{old}} \le \pi_\theta \le c\pi_{\text{old}} \\ 0 & \text{if } \pi_\theta < (1-\epsilon_{\text{low}})\pi_{\text{old}} \quad \text{(Lower Clip)} \\ 0 & \text{if } \pi_\theta > c\pi_{\text{old}} \quad \text{(Dual Clip)} \end{cases} \tag{A6}$$

In the dual-clip region $\pi_\theta > c\pi_{\text{old}}$, the gradient is already zero, so **Eq. A6** sets Cap to zero. This gives the blue regions in Fig. 1(c).

B Training and Evaluation Details

Each baseline and UPAO use identical training setups, hardware allocations, and decoding strategies within each model class; only the objective changes: for tokens with positive advantages, UPAO replaces DAPO's clipped IS term with the self-anchored ratio $\tilde{r}_{i,t}(\theta)$ (Eq. 6), so no clipping applies to them, while tokens with non-positive advantages keep DAPO's lower clip and dual clip (ϵ_{low}, $c = 3$). All experiments use 16 NVIDIA H100 GPUs. Rollout n is the group size G in Eq. 8.

Table A1 and Table A2 give the configurations for Qwen3-14B-Base and for Qwen3-30B-A3B-Base, where UPAO is applied on top of GSPO and keeps its lower clip for non-positive advantages.

Item	**DAPO**	**UPAO**
Prompt / response max	2,048 / 20,480	2,048 / 20,480
Prompts per rollout	512	512
Rollout n	16	16
Prompts per update	32	32
Learning rate	1×10^{-6}	1×10^{-6}
Lower clip ϵ_{low}	0.2	0.2
Upper clip ϵ_{high}	0.28	–
Training rollout temperature	1.0	1.0
Training rollout Top-p	1.0	1.0
Validation Top-p	0.7	0.7
Top-k	-1	-1
Max generation tokens	20,480	20,480

Table A1: Training, algorithmic, and decoding hyperparameters for token-level optimization (DAPO and UPAO on Qwen3-14B-Base).

Item	**GSPO**	**UPAO**
Prompt / response max	2,048 / 20,480	2,048 / 20,480
Prompts per rollout	256	256
Rollout n	16	16
Prompts per update	32	32
Learning rate	1×10^{-6}	1×10^{-6}
Lower clip ϵ_{low}	3×10^{-4}	3×10^{-4}
Upper clip ϵ_{high}	4×10^{-4}	–
Training rollout temperature	1.0	1.0
Training rollout Top-p	1.0	1.0
Validation Top-p	0.7	0.7
Top-k	-1	-1
Max generation tokens	20,480	20,480

Table A2: Training, algorithmic, and decoding hyperparameters for GSPO and UPAO on Qwen3-30B-A3B-Base, where UPAO is applied on top of GSPO.

Table A3 summarizes the Qwen3-8B and Olmo-3-7B-Instruct-SFT comparison on MATH (Levels 3–5); all other baselines use the same settings, and only objective-specific hyperparameters follow their original papers.

Table A4 gives the multimodal configuration for Qwen3-VL-8B-Instruct on Geometry3K.

Table A5 summarizes the policy-staleness study. DAPO and UPAO use identical settings in every case. Each rollout batch receives one optimization pass; we generate eight responses per prompt, use 32 prompts per update, and perform 1,200 updates total. Let B_{roll} be prompts per rollout, B_{upd} prompts per update, N_{resp} responses per rollout, and m updates per rollout. They satisfy

$$N_{\text{resp}} = 8B_{\text{roll}}, \qquad m = \frac{B_{\text{roll}}}{B_{\text{upd}}}, \qquad \frac{N_{\text{resp}}}{m} = 8B_{\text{upd}} = 256, \tag{A7}$$

Increasing B_{roll} from 64 to 512 at fixed $B_{\text{upd}} = 32$ raises m from 2 to 16; examples used by later updates therefore come from an increasingly stale behavior policy. The resulting gradient weights and retained-signal formulas are given in **Eq. A8** and **Eq. A10**.

Item	DAPO	UPAO
Prompt / response max	1,024 / 3,072	1,024 / 3,072
Prompts per rollout	128	128
Rollout n	8	8
Prompts per update	32	32
Learning rate	1×10^{-6}	1×10^{-6}
Lower clip ϵ_{low}	0.2	0.2
Upper clip ϵ_{high}	0.28	–
Training rollout temperature	1.0	1.0
Training rollout Top-p	1.0	1.0
Validation Top-p	0.7	0.7
Top-k	-1	-1
Max generation tokens	3,072	3,072

Table A3: Training, algorithmic, and decoding hyperparameters for DAPO and UPAO in the comprehensive baseline comparison on Qwen3-8B trained on MATH (Levels 3–5).

Item	DAPO	UPAO
Prompt / response max	1,024 / 2,048	1,024 / 2,048
Prompts per rollout	512	512
Rollout n	5	5
Prompts per update	128	128
Learning rate	1×10^{-6}	1×10^{-6}
Lower clip ϵ_{low}	0.2	0.2
Upper clip ϵ_{high}	0.28	–
Training rollout temperature	1.0	1.0
Training rollout Top-p	1.0	1.0
Validation Top-p	0.7	0.7
Top-k	-1	-1
Max generation tokens	2,048	2,048

Table A4: Training, algorithmic, and decoding hyperparameters for multimodal reasoning (DAPO and UPAO on Qwen3-VL-8B-Instruct trained on Geometry3K).

Updates per Rollout m	Prompts per Rollout B_{roll}	Responses per Rollout N_{resp}	Prompts per Update B_{upd}
2	64	512	32
4	128	1,024	32
8	256	2,048	32
16	512	4,096	32

Table A5: Rollout and optimization configurations used to control policy staleness in the seven-benchmark comparison of DAPO and UPAO.

C WHY UPAO RETAINS MORE GRADIENT SIGNAL FOR TOKENS WITH POSITIVE ADVANTAGES UNDER HIGH POLICY STALENESS

This section analyzes one mechanism behind the policy-staleness results in **Sec. 5.2** and Fig. 3(a): how much of the on-policy gradient for tokens with positive advantages each method retains as the policy drifts. We restrict attention to these tokens because DAPO and UPAO use the same objective for tokens with non-positive advantages.

C.1 SETUP

A rollout batch generated by $\pi_{\text{old}} = \pi_{\theta_0}$ is divided into m update batches with the same number of prompts and consumed in a single pass through m sequential optimizer updates $\theta_0 \to \cdots \to \theta_m$ (Table A5). A token in update batch $j \in \{0, \ldots, m-1\}$ is therefore evaluated at θ_j, after j updates. Equal batch sizes make the drift level j uniform on $\{0, \ldots, m-1\}$, and shuffling the prompts before the split makes j independent of the advantage and of π_{old}.

For a token o_t with prefix $s_t = (q, o_{<t})$ evaluated at drift level j, let $r_j = \pi_{\theta_j}(o_t \mid s_t)/\pi_{\text{old}}(o_t \mid s_t)$ and $u = 1 + \epsilon_{\text{high}}$. For $\hat{A} > 0$, the per-token gradients are $g_{\text{DAPO}} = \hat{A}\, r_j \mathbf{1}\{r_j \le u\} \nabla_\theta \log \pi_{\theta_j}$ (Eq. 4) and $g_{\text{UPAO}} = \hat{A}\, \nabla_\theta \log \pi_{\theta_j}$ (Eq. 7). Rewriting the expectation over $o_t \sim \pi_{\text{old}}(\cdot \mid s_t)$ as one over $o_t \sim \pi_{\theta_j}(\cdot \mid s_t)$, with the prefix distribution and $\hat{A}$ held fixed as in the token-level surrogate, gives

$$\mathbb{E}_{\pi_{\text{old}}}[g_\bullet] = \mathbb{E}_{\pi_{\theta_j}}\left[w_\bullet(r_j)\, \hat{A}\, \nabla_\theta \log \pi_{\theta_j} \right], \qquad w_{\text{DAPO}}(r) = \mathbf{1}\{r \le u\}, \qquad w_{\text{UPAO}}(r) = \frac{1}{r}, \tag{A8}$$

for $\bullet \in \{\text{DAPO}, \text{UPAO}\}$. Relative to the on-policy gradient, DAPO keeps the full weight whenever $r \le u$, including $r < 1 - \epsilon_{\text{low}}$, and none otherwise, while UPAO's weight $1/r$ is a discount for $r > 1$ and an amplification for $r < 1$. At $j = 0$, π_{old} is evaluated at θ_0, so $r = 1$ and the two methods coincide.

C.2 RETAINED SIGNAL UNDER POLICY DRIFT

Definition A1 (Retained signal)**.** By Eq. A8, the gradient of method $\bullet$ on the tokens evaluated at drift level j equals the on-policy gradient with each token reweighted by $w_\bullet(r_j)$. The retained signal of $\bullet$ is the average weight over the m updates of one rollout,

$$\rho_\bullet(m) = \frac{1}{m} \sum_{j=0}^{m-1} \mathbb{E}[w_\bullet(r_j)], \tag{A9}$$

where the expectation is over the sampled tokens with positive advantages in update batch j, i.e., under $o_t \sim \pi_{\text{old}}(\cdot \mid s_t)$.

Assumption A2 (Deterministic homogeneous drift)**.** There is a constant $\delta > 0$, the same for DAPO and UPAO, such that every optimizer update raises the log-probability of every token with a positive advantage by exactly δ, so that $r_j = e^{j\delta}$ for $j = 0, \ldots, m-1$. In the single-pass setting, a token in update batch j has not been used for any update when it is evaluated, so $\delta > 0$ is an assumption about how the updates on other prompts generalize to it. Because $\pi_{\theta_j} \le 1$, the assumption is feasible only for tokens with $\pi_{\text{old}} \le e^{-(m-1)\delta}$; tokens with $\pi_{\text{old}} > 1/u$ satisfy $r \le 1/\pi_{\text{old}} < u$, never reach the clipping boundary, and are always retained by DAPO.

Proposition A3 (Retained signal under deterministic drift)**.** *Under Assumption A2, let* $m_u = \lfloor \log u/\delta \rfloor + 1$ *be the number of drift levels inside the clipping boundary. Then*

$$\rho_{\text{DAPO}}(m) = \frac{\min(m, m_u)}{m}, \qquad \rho_{\text{UPAO}}(m) = \frac{1 - e^{-m\delta}}{m\,(1 - e^{-\delta})}. \tag{A10}$$

The ratio $R(m) = \rho_{\text{UPAO}}(m)/\rho_{\text{DAPO}}(m)$ *satisfies* $R(m) \le 1$ *and is non-increasing for* $m \le m_u$*, whereas for* $m \ge m_u$ *it equals* $(1 - e^{-m\delta})\, L(\delta)$ *and increases to*

$$L(\delta) = \lim_{m \to \infty} R(m) = \frac{1}{(1 - e^{-\delta}) m_u}, \qquad \frac{1}{\log u + \delta} \le L(\delta) < \frac{e^\delta}{\log u}. \tag{A11}$$

In particular, $\lim_{\delta \to 0} L(\delta) = 1/\log(1 + \epsilon_{\text{high}})$*. If* $\log u < 1/2$*, then* $L(\delta) > 1$ *for every* $\delta > 0$*.*

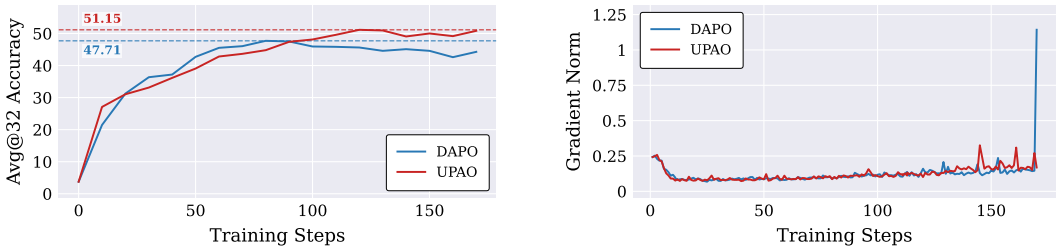
Proof. Under Assumption A2, Eq. A8 gives level j the weights $\mathbf{1}\{j\delta \le \log u\}$ and $e^{-j\delta}$. The levels with $j\delta \le \log u$ are $j = 0, \dots, m_u - 1$, and summing the finite geometric series $\sum_{j=0}^{m-1} e^{-j\delta}$ gives Eq. A10. For $m \le m_u$, $\rho_{\text{DAPO}}(m) = 1$, so $R(m) = \rho_{\text{UPAO}}(m)$ is the running average of the decreasing weights $e^{-j\delta}$, which starts at $R(1) = 1$ and does not increase. For $m \ge m_u$, $\rho_{\text{DAPO}}(m) = m_u/m$, so $R(m) = (1 - e^{-m\delta})/[(1 - e^{-\delta})m_u]$, which increases in m and converges to $L(\delta)$. The definition of m_u gives $\log u/\delta < m_u \le \log u/\delta + 1$, and $\delta e^{-\delta} \le 1 - e^{-\delta} \le \delta$. Hence $(1 - e^{-\delta})m_u \le \delta(\log u/\delta + 1) = \log u + \delta$ and $(1 - e^{-\delta})m_u > \delta e^{-\delta} \log u/\delta = e^{-\delta} \log u$, which gives the bracket; letting $\delta \to 0$ in the bracket gives the limit. Finally, suppose $\log u < 1/2$. If $\delta \le \log u$, the bracket gives $L(\delta) \ge 1/(\log u + \delta) \ge 1/(2 \log u) > 1$. If $\delta > \log u$, then $m_u = 1$ and $L(\delta) = 1/(1 - e^{-\delta}) > 1$. □

Both retained signals decay as $\Theta(1/m)$, but their constants differ. For $m \le m_u$ no level is clipped, so DAPO retains all signal and $R(m) \le 1$. For $m \ge m_u$, $R(m)$ increases, exceeds one once $1 - e^{-m\delta} > (1 - e^{-\delta})m_u$, which for small δ occurs at $m \approx -\log(1 - \log u)/\delta \approx 1.15 m_u$, and converges to $L(\delta)$. DAPO uses $\epsilon_{\text{high}} = 0.28$, so $\log u \approx 0.247 < 1/2$, $L(\delta) > 1$ for every δ, and $L(\delta) \to 1/\log(1.28) \approx 4.05$ as $\delta \to 0$; since $R(m) = (1 - e^{-m\delta})\, L(\delta)$ for $m \ge m_u$, this value is approached only when $m\delta \gg 1$.

D EXPLORATION AND STABILITY CURVES

This appendix provides the supplementary training curves for the generalization experiments in Sec. 5.3, in the same order as that section: the larger-model setting first, followed by the MoE and vision-language settings.

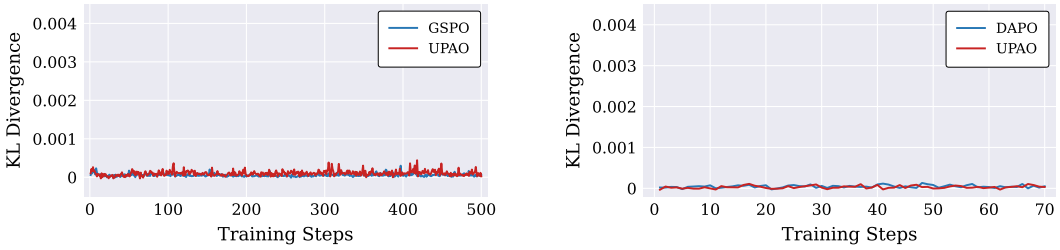
Larger Models. These curves supplement the Qwen3-14B-Base results in the Larger Models paragraph of Sec. 5.3. **Fig. A1(a)** tracks Avg@32 on AIME24 during training: UPAO reaches a peak of 51.15%, compared with 47.71% for DAPO, and maintains a clear advantage later in training, consistent with the Pass@K gains in Fig. 4(a). Fig. A1(b) shows that the gradient norm of UPAO stays comparable to that of DAPO, supporting the main-text claim that UPAO enlarges exploration without compromising training stability.



(a) Avg@32 on AIME24 (b) Gradient norm

Figure A1: Performance and stability results for UPAO on Qwen3-14B-Base trained on DAPO-17K-MATH: (a) Avg@32 on AIME24 and (b) gradient norms of DAPO and UPAO.

MoE Architectures and Vision-Language Models. These curves supplement the MoE Architectures and Vision-Language Models paragraphs of Sec. 5.3, whose accuracy curves are shown in Fig. 5. On Qwen3-30B-A3B-Base trained on DAPO-17K-MATH, the KL divergence of UPAO to the reference model follows a trajectory nearly identical to that of GSPO throughout training (Fig. A2(a)); on Qwen3-VL-8B-Instruct trained on Geometry3K, UPAO likewise tracks DAPO closely (Fig. A2(b)). Thus, UPAO improves accuracy without inducing additional policy drift across architectures and modalities.



(a) UPAO on Qwen3-30B-A3B-Base (b) UPAO on Qwen3-VL-8B-Instruct

Figure A2: Stability curves for the universality experiments: (a) UPAO vs. GSPO on Qwen3-30B-A3B-Base trained on DAPO-17K-MATH and (b) UPAO vs. DAPO on Qwen3-VL-8B-Instruct trained on Geometry3K.

E BROADER IMPACT

UPAO is a generic modification to the policy loss in RLVR that can improve reasoning accuracy and training stability without additional compute. It may therefore benefit scientific, mathematical, and educational applications that rely on stronger reasoning. UPAO inherits standard RLVR risks, including reward hacking, plausible but inaccurate outputs, and biased reward signals. Because it relaxes the trust region for tokens with positive advantages, unreliable rewards may reinforce undesirable behavior, and standard reward-model validation and monitoring remain necessary.

F LIMITATIONS

Our experiments are limited to sparse-reward reasoning tasks in mathematics and visual geometry. UPAO may generalize to settings such as agentic tool use, RLHF with preference rewards, or dense-reward control because it only modifies the importance-sampling ratio for tokens with positive advantages, but this has not yet been verified empirically.