ByteDance | Seed

UP: Unbounded Positive Asymmetric Optimization for Breaking the Exploration-Stability Dilemma

Chongyu Fan[1,2,*,†], **Pengfei Liu**[1], **Jingjia Huang**[1], **Sijia Liu**[2], **Yi Lin**[1,†]

[1]ByteDance Seed, [2]Michigan State University

[*]Work done at ByteDance Seed, [†]Corresponding authors

Abstract

Reinforcement learning (RL) has become the standard paradigm for enhancing the complex reasoning capabilities of large language models (LLMs). To achieve sample efficiency, modern RL frameworks rely on importance sampling (IS). However, these algorithms suffer from an exploration-stability dilemma. Pure IS often leads to catastrophic training instability, while standard clipping mechanisms used to mitigate this instability strictly constrain the policy update budget. By formalizing the concept of **Probability Capacity (Cap),** we reveal that conservative clipping structurally stifles exploration by prematurely truncating the update budget for correct but low-confidence reasoning paths. To break free from these constraints, we propose **Unbounded Positive Asymmetric Optimization (UP)**, a universal and plug-and-play objective. UP theoretically restructures the optimization process by anchoring the policy to its current state via the stop-gradient operator. This asymmetric design unleashes unclipped, stable gradients for positive advantages to maximize exploration, while maintaining standard clipping safeguards for negative advantages to prevent training instability. Furthermore, our formulation readily extends across different optimization granularities, including token-level (GRPO, DAPO) and sequence-level (GSPO) frameworks. Extensive experiments demonstrate that UP enhances exploration capacity and achieves superior reasoning accuracy across diverse RL algorithms (DAPO, GSPO, and GRPO), model architectures (Dense, MoE, and vision-language), and training modalities (language and multimodal), validating UP as a truly universal plug-and-play enhancement for RL-based training.

Date: July 9, 2026
Correspondence: Chongyu Fan at fanchon2@msu.edu, Yi Lin at linyi.james@bytedance.com
Project Page: https://chongyu-fan.netlify.app/posts/up/

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 [4, 9, 27, 35]. However, navigating the vast and intricate search space of reasoning trajectories using supervised fine-tuning alone is often infeasible [16, 38]. To address this challenge, reinforcement learning (RL) has emerged as the standard paradigm [2, 22]. While foundational on-policy methods like REINFORCE [33] offer mathematically stable optimization, they suffer from severe sample inefficiency. Consequently, modern RL frameworks such as Group Relative Policy Optimization (GRPO) [28] and Dynamic sAmpling Policy Optimization (DAPO) [36] have transitioned to multi-step optimization, which aims to improve sample efficiency by utilizing *importance*

sampling (IS) to estimate target distributions from historical policies (π_{old}). More recently, a growing body of work has extended the GRPO/DAPO paradigm along diverse directions, including sequence-level optimization via GSPO [41], critic-free global normalization via REINFORCE++ [11], and asymmetric importance sampling designs such as TOPR [24] and ASPO [32].

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* [6, 18, 20, 37, 42]. 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* [25, 26]. To mitigate this instability, standard algorithms heavily rely on a *clipping* mechanism, which forcibly bounds the policy update within a predefined trust region to preserve original model representations [15, 26].

The clipping mechanism in standard RL algorithms inherently restricts the allowable changes in token probabilities. Although some recent studies have recognized this limitation [6, 8, 31, 39], they have not provided a fundamental resolution. To formally quantify this constraint, we define the **Probability Capacity (Cap)** as the maximum allowable increase in π_θ for tokens with positive advantages, or the maximum allowable decrease for tokens with negative advantages, before the optimization gradient is truncated. For example, DAPO, one of the state-of-the-art algorithms, attempts to mitigate exploration stagnation by utilizing an elevated upper clip bound (ϵ_{high}) for positive advantages. However, as we formally show in Sec. 3, DAPO still remains vulnerable to stifled exploration. By analyzing the Cap, we reveal that adjusting ϵ_{high} only alleviates the issue without fundamentally resolving it. Because the clipping mechanism strictly restricts the Cap for correct but low-confidence logic paths to be linearly proportional to the historical policy (π_{old}), these tokens still receive a severely constrained update budget. Once the policy marginally improves, the gradient is prematurely truncated to zero. This highlights the need to develop a new mathematical foundation to break free from conservative clipping constraints and fully unleash the model's exploration [3, 14, 37].

Motivated by this observation, we ask:

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

Drawing inspiration from the stable, unclipped gradients of REINFORCE, we address *(Q)* through the lens of **Unbounded Positive Asymmetric Optimization (UP)**. Specifically, the treatment of positive advantages focuses on maximizing exploration capacity, coupled with the treatment of negative advantages that acts as a structural safeguard against training instability [24, 32, 43]. We theoretically restructure the importance sampling mechanism by replacing the IS ratio with a self-anchored ratio using the *stop-gradient operator (sg)* [39]. We show that this asymmetric formulation naturally aligns with an unclipped, mathematically stable gradient for correct rollouts. UP, by explicitly bypassing the traditional IS anchor, opens a critical yet underexplored direction for enhancing RL reasoning capabilities without sacrificing optimization stability.

We summarize our **contributions** below.

• We formalize the concept of **Probability Capacity (Cap)** to expose a fundamental dilemma in current RL algorithms: overly conservative clipping constraints structurally stifle exploration for reasoning, whereas overly aggressive updates inevitably lead to training instability.

• We theoretically restructure the optimization process by anchoring the policy to its current state via the stop-gradient operator. Coupled with an in-depth exploration of asymmetric optimization design, we propose **Unbounded Positive Asymmetric Optimization (UP)**, a universal, plug-and-play objective that readily extends across different optimization granularities (e.g., token-level (GRPO, DAPO) and sequence-level (GSPO)).

• We conduct extensive experiments to demonstrate the critical role of the UP formulation in resolving the exploration-stability dilemma. Empirically, UP consistently enhances exploration capacity and achieves superior accuracy across diverse RL algorithms (DAPO, GSPO, and GRPO), model architectures (Dense, MoE, and vision-language), and training modalities (language and multimodal), while preventing training instability across all settings. In addition, against eleven strong RL baselines, including GRPO, Dr. GRPO, CISPO, DPPO, GMPO, GSPO, SAPO, REINFORCE++, RLOO, W-REINFORCE, and ASPO, UP-GRPO attains the best average Pass@1 accuracy across five challenging reasoning benchmarks (AIME24, AMC23, MATH500, Minerva, and OlympiadBench).

2 Related Work

Reinforcement Learning for LLM Reasoning. While REINFORCE [33] established the core policy gradient framework, PPO [26] became the standard RLHF/RLAIF paradigm [2, 22] by improving sample efficiency. However, PPO's auxiliary critic model limits scalability for long-context reasoning. To alleviate this, GRPO [28] omits the critic via group-level relative scoring, driving breakthroughs in models like DeepSeek-R1 [9] and Qwen2.5-Math [34]. Building on GRPO, DAPO [36] introduces decoupled clipping boundaries, while GSPO [41] and related works [17, 21] transition to sequence-level importance sampling (IS) to reduce variance. Despite these advancements, current paradigms still inherit restrictive symmetric clipping mechanisms [18], which we argue structurally stifle exploration.

Importance Sampling and Asymmetric Optimization. To prevent training instability caused by exploding IS ratios, TRPO [25] and PPO introduced heuristic clipping. While recent efforts (e.g., BandPO [15], SAPO [8], GMPO [40], and M2PO [42]) refine these bounds, they all assume the historical policy (π_{old}) must remain in the denominator. This assumption bottlenecks exploration, the most critical challenge in reasoning where golden trajectories are sparse [7, 30, 38], rendering the discovery of these rare paths futile if their gradients are immediately clipped during optimization. Consequently, asymmetric optimization has gained traction to address these conflicting signals (e.g., ASPO [32] and W-REINFORCE [43]). Our UP distinctly advances this paradigm: by introducing a stop-gradient self-anchor ($\text{sg}(\pi_\theta)$), we theoretically restructure the IS ratio to bypass the trust region bottleneck. By combining this Unbounded Positive (UP) formulation for correct rollouts with DAPO's rigorous negative constraints, our method aggressively internalizes long-tail reasoning capabilities without triggering IS instability.

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 Definitions and Formulations: REINFORCE, GRPO, DAPO, and GSPO

REINFORCE. The evolution of RL algorithms is fundamentally driven by the pursuit of sample efficiency and optimization stability. As a foundational approach, the vanilla REINFORCE algorithm optimizes the policy strictly on-policy [33]. Given a query q and a generated response $o_{i,t}$ at step t, the REINFORCE objective is defined as maximizing the expected advantage-weighted log-probability:

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

Taking the derivative of this objective yields a mathematically stable gradient:

$$\nabla_\theta\mathcal{J}_{\text{REINFORCE}}(\theta) = \mathbb{E}_{q\sim\mathcal{Q},o\sim\pi_\theta}\left[\sum_{t=1}^{|o|}\hat{A}\nabla_\theta\log\pi_\theta(o_{i,t}|q,o_{i,<t})\right] \tag{2}$$

GRPO and DAPO. While mathematically stable, REINFORCE suffers from severe sample inefficiency because trajectories must be discarded after a single gradient step. To enable multi-step optimization over the same sampled data, modern frameworks like Group Relative Policy Optimization (GRPO) introduce **Importance Sampling (IS)** [28]. Mathematically, IS is a statistical technique used to estimate the properties of a target distribution (the current policy π_θ) using samples drawn from a distinct proposal distribution (the historical policy π_{old}). This is achieved via the importance sampling ratio:

$$r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\text{old}}(o_{i,t}|q,o_{i,<t})} \tag{3}$$

Despite its sample efficiency, pure unclipped IS suffers from severe optimization instability, as we formally demonstrate in **Sec. 3.2**. To mitigate this instability, GRPO introduces a **clipping** operation. This heuristic regularization technique is designed to restrict the policy update step size by forcibly bounding the importance sampling ratio within a predefined trust region $[1-\epsilon, 1+\epsilon]$. The full GRPO objective is formulated as:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim\mathcal{Q},\{o_i\}_{i=1}^{G}\sim\pi_{\text{old}}} \left[\frac{1}{G}\sum_{i=1}^{G} \frac{1}{|o_i|}\sum_{t=1}^{|o_i|} \left\{ \min\left[r_{i,t}(\theta)\hat{A}_i, \text{clip}(r_{i,t}(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_i \right] - \beta\mathbb{D}_{\text{KL}}(\pi_\theta\|\pi_{\text{ref}}) \right\} \right] \tag{4}$$

where $\hat{A}_i$ is the group-normalized advantage. Building upon this foundation, state-of-the-art methods like Dynamic sAmpling Policy Optimization (DAPO) introduce decoupled clip bounds $(\epsilon_{\text{low}}, \epsilon_{\text{high}})$ while entirely omitting the KL penalty ($\beta = 0$) [36]. Consequently, these methods rely exclusively on clipping to balance optimization stability and enhance exploration capacity, exposing a deep mathematical vulnerability.

GSPO. To address the inherent mismatch between sequence-level rewards and the token-level optimization utilized in GRPO and DAPO, Group Sequence Policy Optimization (GSPO) was recently proposed. GSPO transitions to a sequence-level importance sampling framework [41]. Unlike GRPO, which computes the importance ratio $r_{i,t}(\theta)$ and applies clipping for each individual token, GSPO defines the importance ratio based on the length-normalized likelihood of the entire generated sequence. To prevent gradient variance explosion associated with long reasoning trajectories, GSPO calculates the geometric mean of the token-level ratios. Let $\pi_\theta(o_i|q) = \prod_{t=1}^{|o_i|}\pi_\theta(o_{i,t}|q, o_{i,<t})$ denote the likelihood of the generated response. The length-normalized sequence-level importance ratio $s_i(\theta)$ is defined as:

$$s_i(\theta) = \left(\frac{\pi_\theta(o_i|q)}{\pi_{\text{old}}(o_i|q)}\right)^{\frac{1}{|o_i|}} = \exp\left(\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\left(\log\pi_\theta(o_{i,t}|q, o_{i,<t}) - \log\pi_{\text{old}}(o_{i,t}|q, o_{i,<t})\right)\right) \tag{5}$$

The GSPO objective applies clipping, rewarding, and optimization at the sequence level by directly constraining this aggregated ratio $s_i(\theta)$:

$$\mathcal{J}_{\text{GSPO}}(\theta) = \mathbb{E}_{q\sim\mathcal{Q},\{o_i\}_{i=1}^{G}\sim\pi_{\text{old}}} \left[\frac{1}{G}\sum_{i=1}^{G} \min\left(s_i(\theta)\hat{A}_i, \text{clip}(s_i(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_i \right) \right] \tag{6}$$

3.2 Aggressive Updates Dilemma 1: Importance Sampling Induces Training Instability

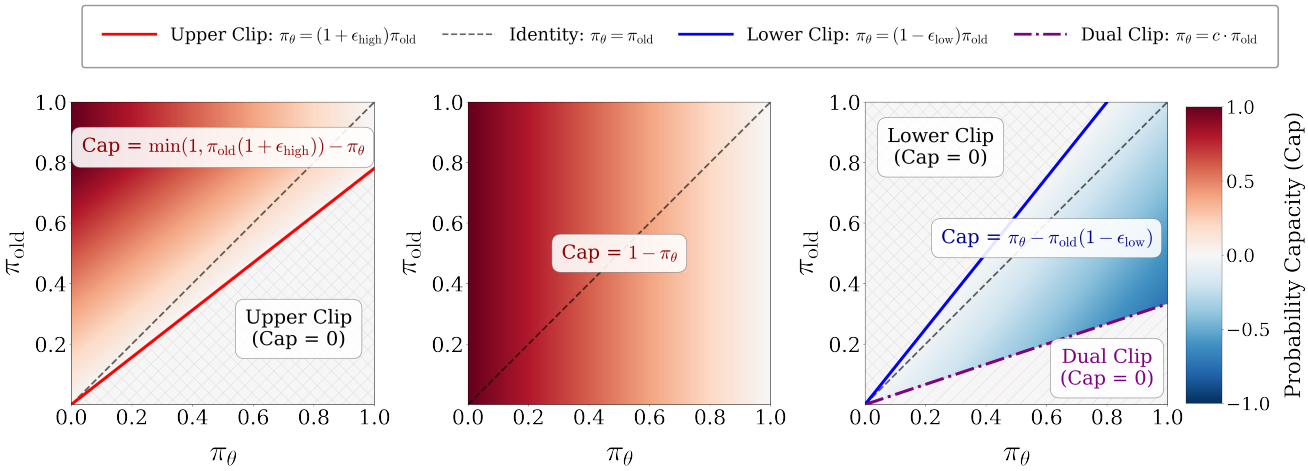
To understand why clipping is considered strictly necessary in the aforementioned methods, we must first formalize the fatal bottleneck introduced by pure IS. The unclipped IS objective seeks to maximize the expected advantage weighted by the importance sampling ratio $r_{i,t}(\theta)$:

$$\mathcal{J}_{\text{IS}}(\theta) = \mathbb{E}_{q\sim\mathcal{Q},o\sim\pi_{\text{old}}} \left[\hat{A} r_{i,t}(\theta) \right] \tag{7}$$

Applying the log-derivative trick, the gradient of this unclipped objective becomes:

$$\nabla_\theta\mathcal{J}_{\text{IS}}(\theta) = \mathbb{E}_{q\sim\mathcal{Q},o\sim\pi_{\text{old}}} \left[\hat{A} r_{i,t}(\theta)\nabla_\theta \log\pi_\theta(o_{i,t}|q, o_{i,<t}) \right] \tag{8}$$

Directly comparing **Eq. 8** to the REINFORCE gradient in **Eq. 2**, the mathematical flaw is evident: the IS gradient is explicitly scaled by $r_{i,t}(\theta)$. For rare, high-reward reasoning paths where the behavior probability $\pi_{\text{old}}(o_{i,t}|q, o_{i,<t})$ is infinitesimally small, $r_{i,t}(\theta)$ explodes as the active policy marginally improves. This disproportionate scaling injects pathological gradients, irrevocably destroying stable representations and inevitably leading to severe **training instability**.



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

Figure 1 Visualization of the probability capacity (Cap) for DAPO and UP-DAPO. 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) UP-DAPO under $\hat{A} > 0$, where the absence of clipping allows all tokens to maintain a corresponding Cap. (c) DAPO and UP-DAPO under $\hat{A} \leq 0$, where ϵ_{low} and c trigger Lower Clip and Dual Clip respectively, leading to Cap $= 0$ in those regions.

3.3 Conservative Constraints Dilemma 2: Clipping Stifles Exploration

To circumvent gradient explosion, standard algorithms like GRPO and DAPO depend heavily on clipping. Specifically, for positive advantages ($\hat{A}_i > 0$) where exploration is most critical for discovering novel and correct reasoning paths, the clipped objective $\mathcal{J}_{\text{IS+CLIP}}$ becomes:

$$\mathcal{J}_{\text{IS+CLIP}}(\theta) = \min\left[r_{i,t}(\theta), 1+\epsilon\right]\hat{A}_i \tag{9}$$

The gradient of this objective with respect to θ is:

$$\nabla_\theta \mathcal{J}_{\text{IS+CLIP}}(\theta) = \begin{cases} \hat{A}_i r_{i,t}(\theta) \nabla_\theta \log \pi_\theta(o_{i,t}|q, o_{i,<t}) & \text{if } r_{i,t}(\theta) \leq 1+\epsilon \\ 0 & \text{if } r_{i,t}(\theta) > 1+\epsilon \end{cases} \tag{10}$$

To formally analyze the consequences of this gradient truncation (**Eq. 10**), we define the **Probability Capacity (Cap)** as the maximum allowable increase (when $\hat{A} > 0$) or decrease (when $\hat{A} \leq 0$) in π_θ in the absolute probability space before the update is blocked. This capacity represents the effective "budget" for policy exploration relative to the reference policy π_{old}.

As illustrated in **Fig. 1(a)**, for tokens with positive advantages ($\hat{A}_{i,t} > 0$) in DAPO, the gradient vanishes entirely the exact moment the probability ratio $r_{i,t}(\theta)$ exceeds the trust region boundary $1+\epsilon_{\text{high}}$. Mapping this constraint to the absolute probability space, the policy is strictly upper-bounded by $\pi_\theta \leq (1+\epsilon_{\text{high}})\pi_{\text{old}}$. Considering that predicted probabilities cannot exceed 1, the maximum achievable probability is $\min\left(1, (1+\epsilon_{\text{high}})\pi_{\text{old}}\right)$.

Consequently, the remaining Probability Capacity for a positive update can be explicitly formulated as a piecewise function:

$$\text{Cap}(\pi_\theta, \pi_{\text{old}}) = \begin{cases} \min\left(1, (1+\epsilon_{\text{high}})\pi_{\text{old}}\right) - \pi_\theta & \text{if } \pi_\theta < (1+\epsilon_{\text{high}})\pi_{\text{old}} \\ 0 & \text{if } \pi_\theta \geq (1+\epsilon_{\text{high}})\pi_{\text{old}} \end{cases} \tag{11}$$

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.

4 Breaking the Dilemma: Unbounded Positive Asymmetric Optimization

This section details how we dismantle the exploration-stability dilemma through two core mathematical innovations: replacing the historical policy π_{old} with the current policy π_θ via the stop-gradient operator, and designing a dynamically routed asymmetric optimization framework. Together, these innovations constitute our **Unbounded Positive Asymmetric Optimization (UP)** methodology.

4.1 Unbounded Formulation for Positive Advantages

To eliminate the pathological gradients injected by π_{old} while retaining the multi-step optimization over the same sampled data, we introduce **Unbounded Positive Asymmetric Optimization (UP)**. Crucially, the unbounded mechanism within UP is specifically designed to be applied exclusively to correct rollouts ($\hat{A} > 0$). For these correct rollouts, we propose replacing the standard importance sampling ratio $r_{i,t}(\theta)$ with a self-anchored modified ratio $\tilde{r}_{i,t}(\theta)$. This is achieved by utilizing the **stop-gradient operator (sg)** to explicitly formulate the unbounded positive component of our framework. This component is denoted as $\mathcal{J}^{+}_{\text{UP}}(\theta)$, where the superscript "+" indicates its restriction to correct rollouts. The objective is formulated as:

$$\mathcal{J}^{+}_{\text{UP}}(\theta) = \mathbb{E}_{q\sim\mathcal{Q},o\sim\pi_{\text{old}}}\left[\sum_{t=1}^{|o|}\hat{A}\tilde{r}_{i,t}(\theta)\right] = \mathbb{E}_{q\sim\mathcal{Q},o\sim\pi_{\text{old}}}\left[\sum_{t=1}^{|o|}\hat{A}\frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\text{sg}(\pi_\theta(o_{i,t}|q,o_{i,<t}))}\right], \quad \text{for } \hat{A}>0 \tag{12}$$

where $\tilde{r}_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\text{sg}(\pi_\theta(o_{i,t}|q,o_{i,<t}))}$.

During backpropagation, the term $\text{sg}(\pi_\theta(o_{i,t}|q,o_{i,<t}))$ functions strictly as a constant scalar. Because the denominator $\text{sg}(\pi_\theta)$ is equal to π_θ in value, taking the derivative of this objective allows us to seamlessly compute the gradient and apply the log-derivative trick ($\nabla x/x = \nabla \log x$) in a single continuous derivation:

$$\begin{aligned}
\nabla_\theta\mathcal{J}^{+}_{\text{UP}}(\theta) &= \mathbb{E}_{q\sim\mathcal{Q},o\sim\pi_{\text{old}}}\left[\sum_{t=1}^{|o|}\hat{A}\frac{1}{\text{sg}(\pi_\theta(o_{i,t}|q,o_{i,<t}))}\nabla_\theta\pi_\theta(o_{i,t}|q,o_{i,<t})\right] \\
&= \mathbb{E}_{q\sim\mathcal{Q},o\sim\pi_{\text{old}}}\left[\sum_{t=1}^{|o|}\hat{A}\frac{1}{\pi_\theta(o_{i,t}|q,o_{i,<t})}\nabla_\theta\pi_\theta(o_{i,t}|q,o_{i,<t})\right] \\
&= \mathbb{E}_{q\sim\mathcal{Q},o\sim\pi_{\text{old}}}\left[\sum_{t=1}^{|o|}\hat{A}\nabla_\theta\log\pi_\theta(o_{i,t}|q,o_{i,<t})\right]
\end{aligned} \tag{13}$$

This substitution yields a profound theoretical conclusion: optimizing this stable, unclipped ratio is mathematically equivalent to maximizing the REINFORCE objective established in **Eq. 1**. By explicitly anchoring the policy to π_θ rather than π_{old}, this formulation completely eradicates the root cause of IS-induced instability. It safely enables unbounded reinforcement for golden reasoning trajectories without triggering the gradient explosion analyzed in **Sec. 3.2**. As shown in **Fig. 1(b)**, the Cap under this formulation becomes fundamentally unconstrained by the historical policy. Defined simply by the maximum allowable increase ($\text{Cap} = 1 - \pi_\theta$), this ensures all tokens maintain a full update budget without clipping.

Building upon this unbounded formulation, we complete the UP framework. We intentionally engineer this asymmetric design to address the divergent mathematical dynamics between correct and wrong rollouts. For correct rollouts, our primary objective is to maximize the exploration capacity and substantially amplify the

reinforcement signal for rare, low-confidence tokens. Therefore, we directly apply the unbounded positive component ($\mathcal{J}_{\text{UP}}^{+}$) to unleash unconstrained reinforcement. Conversely, for wrong rollouts, the advantage is inherently negative, which reverses the direction of the gradient. Applying an unbounded update performs aggressive gradient ascent and destroys the original representation. Consequently, this unbounded mechanism must be strictly prohibited in the negative regime.

4.2 UP-GxPO: Universal Asymmetric Integration

Crucially, the UP framework serves as a universal plugin that can be seamlessly integrated with any Group-based Policy Optimization (GxPO) algorithm. Recognizing these opposing requirements, we apply an asymmetric modification to standard policy objectives by dynamically routing the gradient computation based on the correctness of the rollouts.

In this work, we first instantiate our primary token-level method as **UP-DAPO** by combining the unbounded formulation for positive advantages with the DAPO baseline for negative advantages:

$$\mathcal{J}_{\text{UP-DAPO}}(\theta) = \mathbb{E}_{q\sim\mathcal{Q},\{o_i\}_{i=1}^{G}\sim\pi_{\text{old}}}\left[\frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|} \begin{cases} \hat{A}_{i,t}\log\pi_\theta(o_{i,t}|q,o_{i,<t}) & \text{if } \hat{A}_{i,t}>0 \\ \min\left(r_{i,t}(\theta)\hat{A}_{i,t}, \text{clip}(r_{i,t}(\theta),1-\epsilon_{\text{low}},1+\epsilon_{\text{high}})\hat{A}_{i,t}\right) & \text{if } \hat{A}_{i,t}\le 0 \end{cases}\right], \tag{14}$$

where the objective for wrong rollouts utilizes the standard decoupled clipping mechanism from DAPO to penalize incorrect actions.

This asymmetric design explicitly resolves the central dilemma analyzed in the previous section. For correct rollouts ($\hat{A} > 0$), we completely discard the clipping mechanism, actively overcoming the conservative constraints and the capacity mismatch identified previously. This deliberate unboundedness maximizes the exploration for rare, low-confidence tokens, allowing the model to aggressively reinforce successful long-tail reasoning paths. Conversely, for wrong rollouts ($\hat{A} \le 0$), we retain the standard DAPO clipping mechanism as a critical structural safeguard. We visualize the Cap for both DAPO and UP-DAPO under $\hat{A} \le 0$ in **Fig. 1(c)**. Detailed derivations are provided in **Appendix B**. This ensures we do not execute overly aggressive penalization updates on wrong rollouts, thereby strictly preventing training instability.

The same asymmetric principle applies directly to GRPO [28], another widely used token-level algorithm that employs symmetric clipping (a single ϵ for both bounds) and a sequence-level group-normalized advantage $\hat{A}_i$ shared across all tokens within the same rollout. **UP-GRPO** replaces the clipped update for positive advantages with the unbounded log-policy objective, while retaining the standard GRPO clipping safeguard and KL penalty for negative advantages:

$$\mathcal{J}_{\text{UP-GRPO}}(\theta) = \mathbb{E}_{q\sim\mathcal{Q},\{o_i\}_{i=1}^{G}\sim\pi_{\text{old}}}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|} \begin{cases} \hat{A}_i\log\pi_\theta(o_{i,t}|q,o_{i,<t}) - \beta\mathbb{D}_{\text{KL}}(\pi_\theta\|\pi_{\text{ref}}) & \text{if } \hat{A}_i>0 \\ \min\left(r_{i,t}(\theta)\hat{A}_i, \text{clip}(r_{i,t}(\theta),1-\epsilon,1+\epsilon)\hat{A}_i\right) - \beta\mathbb{D}_{\text{KL}}(\pi_\theta\|\pi_{\text{ref}}) & \text{if } \hat{A}_i\le 0 \end{cases}\right], \tag{15}$$

where $\hat{A}_i$ is the group-normalized advantage shared across all tokens of rollout o_i, and the negative branch retains the original symmetric GRPO clipping and KL penalty to prevent training instability.

Furthermore, to demonstrate its universality across different optimization granularities, our UP framework readily extends to sequence-level algorithms, yielding **UP-GSPO**. Using the length-normalized sequence-level importance ratio $s_i(\theta) = \left(\frac{\pi_\theta(o_i|q)}{\pi_{\text{old}}(o_i|q)}\right)^{\frac{1}{|o_i|}}$, the overall UP-GSPO objective is formulated as:

$$\mathcal{J}_{\text{UP-GSPO}}(\theta) = \mathbb{E}_{q\sim\mathcal{Q},\{o_i\}_{i=1}^{G}\sim\pi_{\text{old}}}\left[\frac{1}{G}\sum_{i=1}^{G} \begin{cases} \hat{A}_i\left(\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\log\pi_\theta(o_{i,t}|q,o_{i,<t})\right) & \text{if } \hat{A}_i>0 \\ \min\left(s_i(\theta)\hat{A}_i,\ \text{clip}(s_i(\theta),1-\epsilon,1+\epsilon)\hat{A}_i\right) & \text{if } \hat{A}_i\le 0 \end{cases}\right]. \tag{16}$$

The exact analytical derivation, which rigorously proves that the positive branch of UP-GSPO mathematically equates to an unclipped, length-normalized REINFORCE gradient, is detailed in **Appendix C**.

5 Experiments

5.1 Experimental Setup

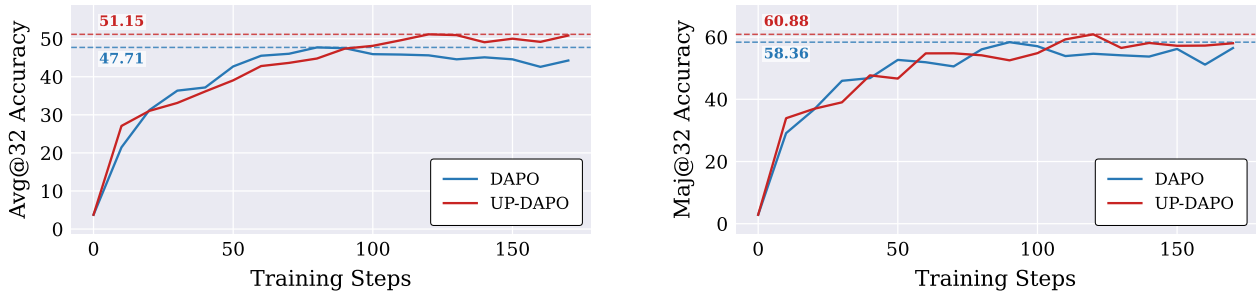
Models. We evaluate the UP framework across diverse model families [35], including dense LLMs in both base (Qwen3-14B-Base) and instruct (Qwen3-8B) versions, an MoE model (Qwen3-30B-A3B-Base), and a vision-language model (Qwen3-VL-8B-Instruct).

Training and Evaluation. We adopt three protocols: (i) following Yu et al. [36], we train on DAPO-17K-MATH and evaluate on AIME24 [5], reporting average accuracy (Avg@32) and majority voting accuracy (Maj@32) over 32 sampled trajectories, together with Best@32 (the probability of generating at least one correct answer within 32 samples); (ii) following Liu et al. [18], we train on MATH (Levels 3-5) [16] and evaluate Pass@1 on AIME24 [5], AMC23, MATH500, Minerva [13], and OlympiadBench [10]; (iii) following Zhao et al. [40], we train on the Geometry3K [19] training set and evaluate Pass@1 on the Geometry3K test set.

Algorithms and Baselines. We instantiate three UP variants under our framework: **UP-GRPO**, **UP-DAPO**, and **UP-GSPO**. We compare against twelve representative RL baselines: DAPO [36], GRPO [28], GMPO [40], ASPO [32], CISPO [4], Dr. GRPO [18], W-REINFORCE [43], REINFORCE++ [11], RLOO [1], DPPO [23], GSPO [41], and SAPO [8].

Implementation Details. Our implementation is based on the verl framework [29], utilizing vLLM for efficient rollout generation and evaluation [12]. Comprehensive details are provided in **Appendix A**.

5.2 Performance, Exploration, and Stability Analysis of UP-DAPO

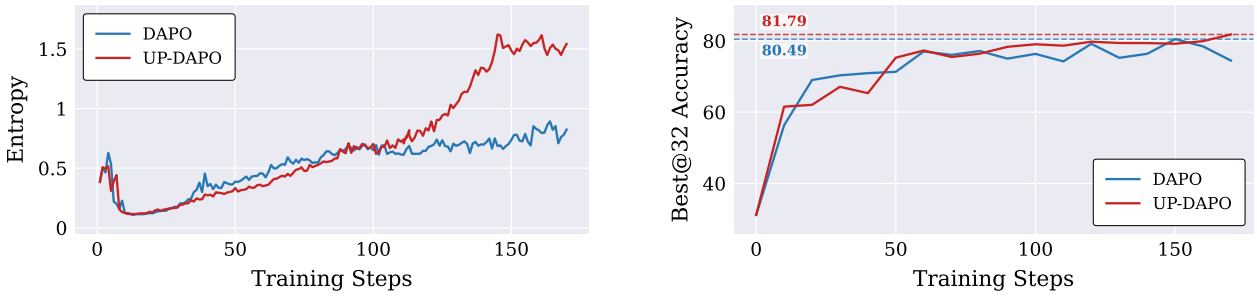


(a) Avg@32 **(b)** Maj@32

Figure 2 Performance comparison of DAPO and UP-DAPO on Qwen3-14B-Base during training: (a) Avg@32 and (b) Maj@32 accuracy on AIME24. Peak performance for each method is denoted by dashed lines and colored labels.

Evaluation on AIME24 during Training. As illustrated in **Fig. 2**, we monitor the performance evolution of DAPO and **UP-DAPO** on the AIME24 evaluation set throughout the training trajectory. The results in Fig. 2(a) and Fig. 2(b) show that UP-DAPO achieves superior performance over the standard DAPO baseline as training progresses. Specifically, UP-DAPO achieves a peak Avg@32 of 51.15, significantly higher than DAPO's 47.71. For Maj@32, UP-DAPO reaches 60.88, surpassing DAPO's 58.36. Notably, the evaluation curves show that UP-DAPO maintains an upward momentum and establishes a clear performance gap in the later stages of optimization, suggesting that the removal of the π_{old} anchor allows for more effective policy optimization in reasoning tasks.

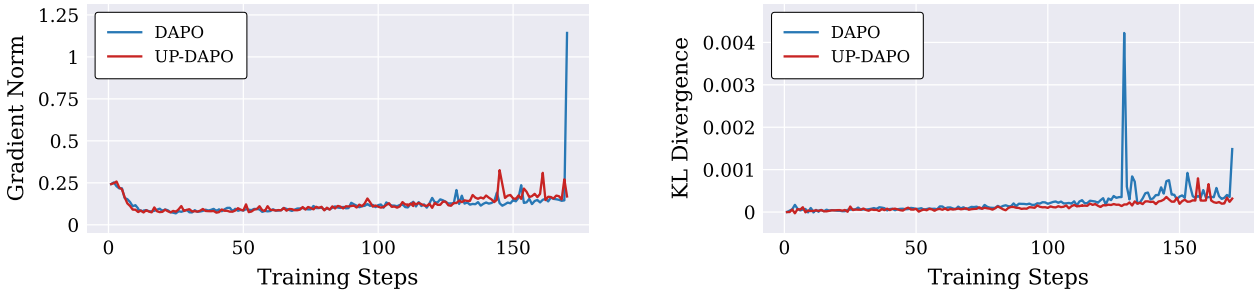
Enhancement of Exploration Capacity. In **Fig. 3(a)**, we present the entropy of generation probabilities for UP-DAPO and DAPO on the DAPO-17K-MATH training set to evaluate their respective exploration capacities. It is evident that UP-DAPO exhibits higher entropy compared to DAPO, which directly correlates to a maximized exploration capacity during the training process. To further validate this, we report the Best@32 accuracy on the AIME24 evaluation set throughout the training trajectory in **Fig. 3(b)**. The results



(a) Entropy **(b)** Best@32

Figure 3 Exploration capacity of DAPO and UP-DAPO on Qwen3-14B-Base during training: (a) entropy of generation probabilities on DAPO-17K-MATH and (b) Best@32 accuracy on AIME24. Peak performance for each method is denoted by dashed lines and colored labels.

show that UP-DAPO achieves a peak Best@32 of 81.79, surpassing DAPO's 80.49. This improvement in the performance upper bound confirms that the increased entropy in UP-DAPO successfully translates into a stronger exploration capability, allowing the model to discover higher-quality reasoning paths.



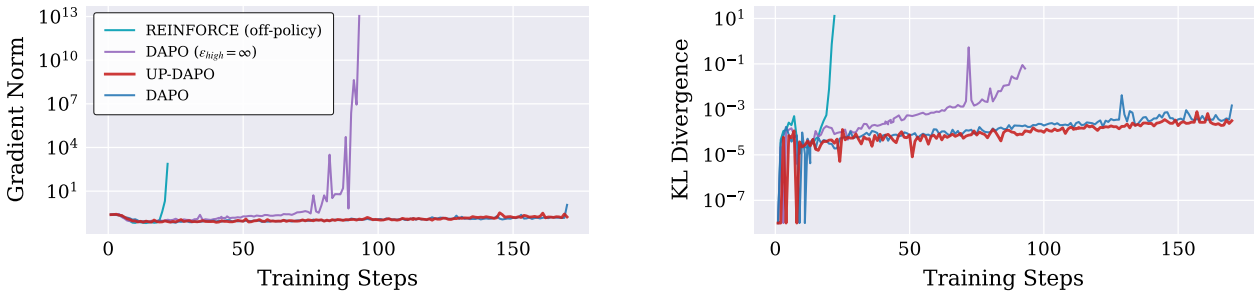
(a) Gradient norm **(b)** KL divergence

Figure 4 Training stability of DAPO and UP-DAPO on Qwen3-14B-Base during training: (a) gradient norm and (b) KL divergence between the active policy and the reference model on DAPO-17K-MATH.

Analysis of Training Stability. In **Fig. 4**, we evaluate the stability of the training process. Despite the increased exploration observed in previous metrics, the gradient norm in **Fig. 4(a)** and the KL divergence in **Fig. 4(b)** of UP-DAPO remain comparable to, or even slightly lower than, those of the standard DAPO baseline. This observation is critical as it demonstrates that although UP-DAPO removes the importance sampling and clipping mechanisms for correct rollouts to maximize exploration, it does so without sacrificing stability or causing the policy to deviate from the reference model. By maintaining these metrics within a stable range, our method effectively prevents training instability while preserving a more expansive exploration capacity.

5.3 Ablation Studies of UP-DAPO

Necessity of Self-Anchored Ratio. We first examine whether the standard DAPO framework can achieve comparable exploration by simply relaxing its constraints within the original importance sampling ratio $r_{i,t}(\theta)$. This experiment is designed to demonstrate the necessity of our self-anchored ratio $\tilde{r}_{i,t}(\theta)$ in **Eq. 12**. We implement a modified DAPO baseline that sets the upper clip bound to infinity ($\epsilon_{\text{high}} = \infty$). As shown by the purple curves in **Fig. 5**, while this modification initially remains stable, it induces severe training instability beyond 80 steps. The gradient norm explodes to 10^{13}, accompanied by a surge in KL divergence. This instability demonstrates that merely removing the upper clip bound while still anchoring the ratio to π_{old} is fundamentally insufficient for safe exploration. Our findings highlight the critical role of the stop-gradient operator: by replacing π_{old} with $\text{sg}(\pi_\theta)$ to formulate the self-anchored modified ratio $\tilde{r}_{i,t}(\theta)$, we effectively eliminate the root cause of instability and safely enable unbounded optimization.



(a) Gradient norm **(b)** KL divergence

Figure 5 Training stability of DAPO, DAPO ($\epsilon_{\text{high}} = \infty$), REINFORCE (off-policy), and UP-DAPO on Qwen3-14B-Base during training: (a) gradient norm and (b) KL divergence between the active policy and the reference model on DAPO-17K-MATH.

Necessity of Asymmetric Objective Design. We further investigate the structural necessity of the asymmetric mechanism in **Eq. 14** by evaluating a symmetric, unbounded baseline. This variant adopts an off-policy REINFORCE-style formulation that applies unbounded updates to both positive and negative advantages, effectively removing all clipping constraints. As shown by the cyan curves in Fig. 5, applying the unbounded mechanism to wrong rollouts leads to immediate and catastrophic training instability within the first 25 training steps. Both the gradient norm and KL divergence exhibit an uncontrolled vertical surge. This confirms that while the unbounded formulation is beneficial for correct rollouts, DAPO's clipping remains a strictly indispensable safeguard for wrong rollouts to prevent training instability.

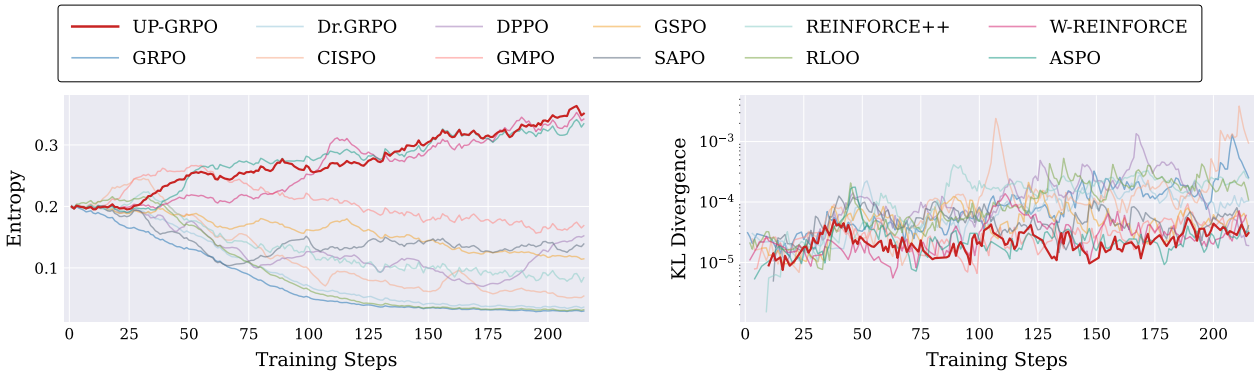
5.4 Comparison with Other RL Baselines

Table 1 Performance of UP-GRPO against eleven RL baselines on AIME24, AMC23, MATH500, Minerva and OlympiadBench. All algorithms are trained on MATH (Levels 3-5) using Qwen3-8B following the protocol of Liu et al. [18]. Performance is evaluated by Pass@1 accuracy (%) on each benchmark and the **Average** across the five benchmarks. The best score in each column is in **bold**.

Method	AIME24	AMC23	MATH500	Minerva	OlympiadBench	Average
GRPO [28]	35.73	75.00	86.00	30.88	51.34	55.79
Dr. GRPO [18]	33.33	85.00	85.80	30.15	51.19	57.09
CISPO [4]	38.02	**87.50**	86.60	29.04	55.65	59.36
DPPO [23]	40.10	**87.50**	86.20	30.51	53.27	59.52
GMPO [40]	37.50	**87.50**	87.00	**31.25**	55.06	59.66
GSPO [41]	40.52	85.00	88.20	**31.25**	55.80	60.15
SAPO [8]	39.90	82.50	87.20	30.88	55.65	59.23
REINFORCE++ [11]	20.52	62.50	78.80	29.78	42.26	46.77
RLOO [1]	31.67	80.00	85.20	28.68	50.60	55.23
W-REINFORCE [43]	35.52	80.00	85.80	30.15	53.27	56.95
ASPO [32]	37.50	85.00	87.60	29.78	**58.48**	59.67
UP-GRPO (Ours)	**41.04**	**87.50**	**88.40**	**31.25**	58.33	**61.31**

Overall Performance. We benchmark **UP-GRPO** against eleven representative RL baselines under the unified protocol of Liu et al. [18], training on MATH (Levels 3-5) with Qwen3-8B and evaluating Pass@1 accuracy across five reasoning benchmarks. As reported in **Table 1**, UP-GRPO attains the best average accuracy of 61.31%, surpassing all competing baselines including the strongest prior method GSPO (60.15%) by an absolute margin of 1.16%. Beyond the average, UP-GRPO ranks first or ties for first on four out of the five individual benchmarks, achieving 41.04% on AIME24, 87.50% on AMC23, 88.40% on MATH500, and 31.25% on Minerva, while remaining highly competitive on OlympiadBench (58.33%, second only to ASPO's 58.48%).

These consistent gains across benchmarks of varying difficulty demonstrate that the unbounded positive update yields a broadly effective, rather than benchmark-specific, improvement over the GxPO family.

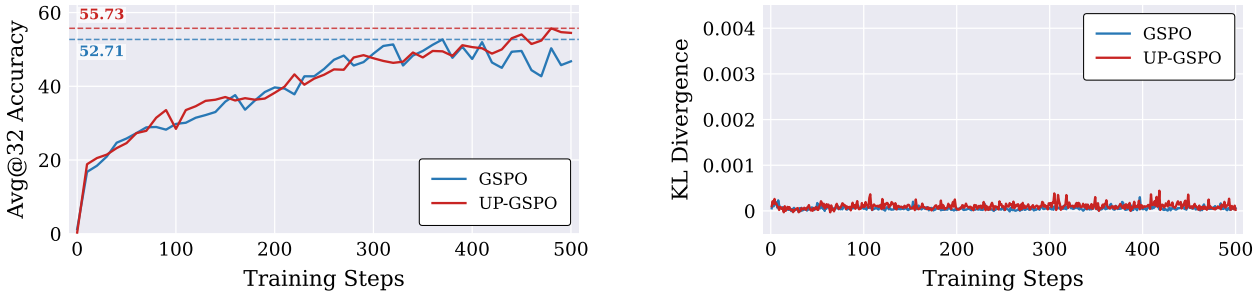


(a) Entropy **(b)** KL divergence

Figure 6 Exploration capacity and training stability of UP-GRPO against eleven 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).

Exploration Capacity and Training Stability. To understand the source of these gains, we further examine the exploration capacity (measured by policy entropy) and the training stability (measured by KL divergence between the active policy and the reference model) of all twelve methods throughout training. **Fig. 6(a)** reveals a clear bifurcation among the baselines. The majority of methods, including GRPO, Dr. GRPO, CISPO, DPPO, GMPO, GSPO, SAPO, REINFORCE++, and RLOO, exhibit a steady entropy collapse as training proceeds, indicating that their policies progressively lose the ability to explore alternative reasoning trajectories. In contrast, UP-GRPO maintains a consistently rising entropy throughout training, confirming that the unbounded positive update preserves and even amplifies exploration capacity. Notably, the only other methods that avoid entropy collapse are W-REINFORCE and ASPO, both of which—like UP-GRPO—adopt asymmetric treatments of correct and wrong rollouts. This shared behavior reaffirms that asymmetric optimization is a necessary structural ingredient for sustaining exploration in long-horizon RL training. **Fig. 6(b)** further shows that UP-GRPO's KL divergence remains among the lowest across the entire training trajectory, indicating that the active policy stays close to the reference model and thereby ensures stable optimization. Taken together, these results show that UP-GRPO simultaneously achieves the strongest exploration capacity and one of the most stable training dynamics, providing a principled explanation for its superior downstream performance.

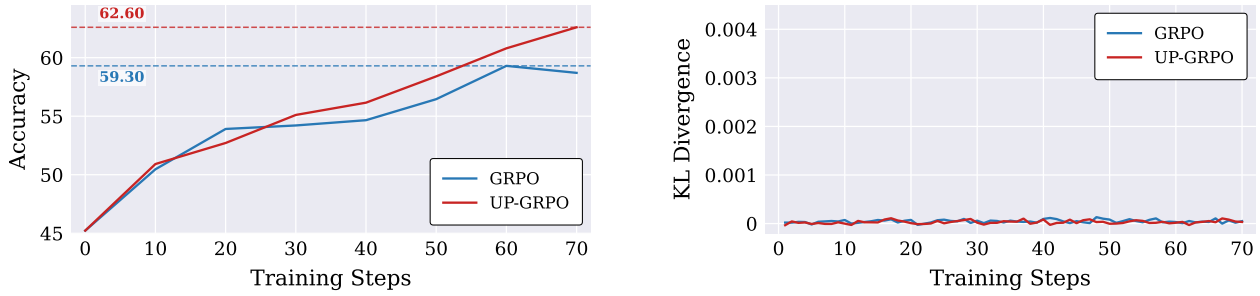
5.5 Universality of UP across Algorithms, Architectures, and Modalities



(a) Avg@32 **(b)** KL divergence

Figure 7 Performance and stability comparison of GSPO and UP-GSPO on Qwen3-30B-A3B-Base during training: (a) Avg@32 accuracy on AIME24 and (b) KL divergence between the active policy and the reference model on DAPO-17K-MATH. Peak performance for each method is denoted by dashed lines and colored labels.

Universality across GxPO Variants and Model Architectures. To validate that the UP framework serves as a universal plugin across diverse GxPO algorithms and model architectures, we evaluate **UP-GSPO** on the Qwen3-30B-A3B-Base MoE model. This experiment tests the framework's adaptability to the architectural shift from Dense to MoE. As illustrated in **Fig. 7**, UP-GSPO consistently achieves superior performance compared to the standard GSPO baseline. Specifically, UP-GSPO reaches a peak Avg@32 accuracy of 55.73%, providing an absolute improvement of 3.02% over GSPO's 52.71%. Crucially, the KL divergence of UP-GSPO remains nearly identical to the baseline throughout the training process. This demonstrates that our design in **Eq. 16** effectively resolves the exploration-stability dilemma, regardless of the optimization granularity (from token-level DAPO to sequence-level GSPO) or the model architecture (from a dense model to MoE). These results confirm that UP is a robust enhancement, capable of safely increasing exploration for correct reasoning trajectories while preventing training instability across the broader GxPO family and diverse architectures.



(a) Accuracy **(b)** KL divergence

Figure 8 Performance and stability comparison of GRPO and UP-GRPO on Qwen3-VL-8B-Instruct during training: (a) accuracy on Geometry3K test set and (b) KL divergence between the active policy and the reference model. Peak performance for each method is denoted by dashed lines and colored labels.

Generalization to Multimodal Setting and New Training Data. To further validate that UP is both modality-agnostic and training-data-agnostic, we evaluate **UP-GRPO** (formulated in **Eq. 15**) against the standard GRPO baseline using the Qwen3-VL-8B-Instruct vision-language model trained on the Geometry3K [19] visual reasoning dataset. This experiment extends our validation beyond language reasoning to multimodal geometric problem solving, where the model must jointly process visual diagrams and textual descriptions. As illustrated in **Fig. 8(a)**, UP-GRPO consistently achieves superior accuracy over GRPO throughout the training process. Specifically, UP-GRPO reaches a peak accuracy of 62.60%, representing an absolute improvement of 3.30% over GRPO's 59.30%. As shown in **Fig. 8(b)**, the KL divergence of UP-GRPO remains nearly identical to that of GRPO, confirming that the unbounded positive update does not induce any additional instability even in the multimodal training regime. These results demonstrate that UP generalizes seamlessly across modalities and datasets, establishing it as a truly universal plug-and-play objective for RL-based training of both language and vision-language models.

6 Conclusion

To overcome the inherent exploration-stability dilemma in IS-based reinforcement learning, we have explored the role of Unbounded Positive Asymmetric Optimization (UP) in maximizing exploration capacity and established its universal applicability across both token-level and sequence-level frameworks. Through the formalization of Probability Capacity (Cap), we have demonstrated how anchoring the policy via the stop-gradient operator mathematically bypasses conservative clipping constraints for positive advantages. Extensive experiments confirm that UP significantly improves reasoning performance across diverse RL algorithms (DAPO, GSPO, and GRPO), model architectures (Dense, MoE, and vision-language), and training modalities (language and multimodal), with our asymmetric design serving as a highly effective and universal paradigm for unleashing exploration while strictly preventing training instability.

References

[1] 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), pages 12248–12267, 2024.

[2] 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.

[3] Hongye Cao, Zhixin Bai, Ziyue Peng, Boyan Wang, Tianpei Yang, Jing Huo, Yuyao Zhang, and Yang Gao. Efficient reinforcement learning with semantic and token entropy for llm reasoning. arXiv preprint arXiv:2512.04359, 2025.

[4] 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.

[5] MAA Codeforces. American invitational mathematics examination-aime 2024, 2024.

[6] Chongyu Fan, Gaowen Liu, Mingyi Hong, Ramana Rao Kompella, and Sijia Liu. Rethinking muon beyond pretraining: Spectral failures and high-pass remedies for vla and rlvr. arXiv preprint arXiv:2605.19282, 2026.

[7] 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.

[8] 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.

[9] 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.

[10] 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), pages 3828–3850, 2024.

[11] 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.

[12] 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, pages 611–626, 2023.

[13] 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.

[14] Pengyi Li, Elizaveta Goncharova, Andrey Kuznetsov, and Ivan Oseledets. Back to basics: Revisiting exploration in reinforcement learning for llm reasoning via generative probabilities. arXiv preprint arXiv:2602.05281, 2026.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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, pages 6774–6786, 2021.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

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

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

[27] Bytedance Seed. Seed2. 0 model card: Towards intelligence frontier for real-world complexity. arXiv preprint arXiv:2607.00248, 2026.

[28] 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.

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

[30] 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.

[31] 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.

[32] 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.

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

[34] An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, et al. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. arXiv preprint arXiv:2409.12122, 2024.

[35] 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.

[36] 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.

[37] 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.

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

[39] 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.

[40] 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.

[41] 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, 2025.

[42] 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, 2025.

[43] 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

A Training and Evaluation Details

In this section, we provide the comprehensive hyperparameter configurations used for training and evaluating. To ensure a strictly fair and rigorous empirical comparison, the baseline algorithms and their corresponding UP variants share identical foundational training setups, hardware allocations, and decoding strategies within their respective model classes. The only variable altered between a baseline and its UP counterpart is the specific mathematical optimization objective introduced in the main text.

Table A1 outlines the core algorithmic and optimization hyperparameters. Crucially, for both UP variants, these clipping bounds are applied asymmetrically: they function strictly as a structural safeguard for negative advantages, leaving positive advantages completely unconstrained. Consequently, the UP variants elegantly eliminate the need for the upper clip hyperparameter ϵ_{high}. **Table A2** summarizes the decoding configurations for the rollout generation phase.

Table A1 Training and algorithmic hyperparameters for token-level (DAPO and UP-DAPO on Qwen3-14B-Base) and sequence-level (GSPO and UP-GSPO on Qwen3-30B-A3B-Base) optimization.

	Dense 14B		MoE 30B	
Item	**DAPO**	**UP-DAPO**	**GSPO**	**UP-GSPO**
Prompt / Response max	2,048 / 20,480	2,048 / 20,480	2,048 / 20,480	2,048 / 20,480
Global batch size (prompts)	512	512	256	256
Rollout n	16	16	16	16
PPO mini-batch size	32	32	32	32
Learning rate	1×10^{-6}	1×10^{-6}	1×10^{-6}	1×10^{-6}
Lower Clip ϵ_{low}	0.2	0.2	3×10^{-4}	3×10^{-4}
Upper Clip ϵ_{high}	0.28	–	4×10^{-4}	–

Table A2 Decoding and rollout hyperparameters for token-level (DAPO and UP-DAPO on Qwen3-14B-Base) and sequence-level (GSPO and UP-GSPO on Qwen3-30B-A3B-Base) training.

Item	**DAPO / UP-DAPO**	**GSPO / UP-GSPO**
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 A3 and **Table A4** provide the training and decoding configurations used by GRPO and UP-GRPO in the comprehensive baseline comparison on Qwen3-8B trained on MATH (level 3-5). The two methods share identical configurations; the only difference is that UP-GRPO removes the upper clipping bound for positive advantages, eliminating the need for the upper clip hyperparameter ϵ_{high}. For the remaining baselines, we keep the training and evaluation setting (*i.e.*, backbone model, training corpus, evaluation benchmarks, batch size, generation budget, and decoding configurations) strictly identical to that of GRPO and UP-GRPO. Only the algorithm-specific hyperparameters in the policy-loss objective are adopted from the original papers.

Table A5 and **Table A6** provide the training and decoding configurations for the multimodal experiment using Qwen3-VL-8B-Instruct on Geometry3K. Both GRPO and UP-GRPO share identical configurations; the only difference is that UP-GRPO removes the upper clipping bound for positive advantages, eliminating the need for the upper clip hyperparameter ϵ_{high}.

Table A3 Training and algorithmic hyperparameters for GRPO and UP-GRPO in the comprehensive baseline comparison on Qwen3-8B trained on MATH (level 3-5).

Item	GRPO	UP-GRPO
Prompt / Response max	1,024 / 3,072	1,024 / 3,072
Global batch size (prompts)	128	128
Rollout n	8	8
PPO mini-batch size	32	32
Learning rate	1×10^{-6}	1×10^{-6}
Lower Clip ϵ_{low}	0.2	0.2
Upper Clip ϵ_{high}	0.2	–

Table A4 Decoding and rollout hyperparameters for GRPO and UP-GRPO in the comprehensive baseline comparison on Qwen3-8B trained on MATH (level 3-5).

Item	GRPO / UP-GRPO
Training rollout temperature	1.0
Training rollout Top-p	1.0
Validation Top-p	0.7
Top-k	-1
Max generation tokens	3,072

Table A5 Training and algorithmic hyperparameters for multimodal reasoning (GRPO and UP-GRPO on Qwen3-VL-8B-Instruct trained on Geometry3K).

Item	GRPO	UP-GRPO
Prompt / Response max	1,024 / 2,048	1,024 / 2,048
Global batch size (prompts)	512	512
Rollout n	5	5
PPO mini-batch size	128	128
Learning rate	1×10^{-6}	1×10^{-6}
Lower Clip ϵ_{low}	0.2	0.2
Upper Clip ϵ_{high}	0.28	–

Table A6 Decoding and rollout hyperparameters for multimodal training (GRPO and UP-GRPO on Qwen3-VL-8B-Instruct trained on Geometry3K).

Item	GRPO / UP-GRPO
Training rollout temperature	1.0
Training rollout Top-p	1.0
Validation Top-p	0.7
Top-k	-1
Max generation tokens	2,048

B Derivation of Probability Capacity for Negative Advantages

In this section, we derive the Probability Capacity (Cap) for the negative advantage regime ($\hat{A} \leq 0$) as analyzed in Sec. 4 and visualized in Fig. 1(c).

For tokens with non-positive advantages, both DAPO and UP-DAPO employ a decoupled clipping mechanism to prevent excessive penalization and training instability. The clipped objective $\mathcal{J}_{\text{neg}}$ is defined as:

$$\mathcal{J}_{\text{neg}}(\theta) = \min\left(r_{i,t}(\theta)\hat{A}, \text{clip}(r_{i,t}(\theta), 1-\epsilon_{\text{low}}, 1+\epsilon_{\text{high}})\hat{A}\right) \tag{A1}$$

An additional Dual Clip constraint c is often introduced to bound the ratio from above when the advantage is negative. Combining these constraints, the gradient $\nabla_\theta \mathcal{J}_{\text{neg}}$ remains non-zero only when the probability ratio $r_{i,t}(\theta)$ stays within the effective optimization window:

$$(1-\epsilon_{\text{low}}) \leq r_{i,t}(\theta) \leq c \tag{A2}$$

where $1-\epsilon_{\text{low}}$ is the lower trust-region boundary and c is the dual-clip threshold.

To derive the Capacity, we map the constraints in Eq. A2 to the absolute probability π_θ. By multiplying the historical policy π_{old}, we identify two critical boundaries:

1. **Lower Clip:** $\pi_{\text{lower}} = (1-\epsilon_{\text{low}})\pi_{\text{old}}$. If $\pi_\theta < \pi_{\text{lower}}$, the gradient is nullified to prevent the probability from dropping too low (over-penalization).
2. **Dual Clip:** $\pi_{\text{upper}} = c \cdot \pi_{\text{old}}$. If $\pi_\theta > \pi_{\text{upper}}$, the gradient vanishes to prevent the policy from moving further away from the reference when the action is already deemed "wrong."

In the negative regime, the goal of optimization is to *decrease* the probability of the token. Thus, the Capacity represents the maximum allowable decrease before the policy hits the lower clip boundary.

Based on the boundaries derived above, the Probability Capacity for $\hat{A} \leq 0$ is formulated as the following piecewise function:

$$\text{Cap}(\pi_\theta, \pi_{\text{old}}) = \begin{cases} \pi_\theta - (1-\epsilon_{\text{low}})\pi_{\text{old}} & \text{if } (1-\epsilon_{\text{low}})\pi_{\text{old}} \leq \pi_\theta \leq c \cdot \pi_{\text{old}} \\ 0 & \text{if } \pi_\theta < (1-\epsilon_{\text{low}})\pi_{\text{old}} \quad \text{(Lower Clip)} \\ 0 & \text{if } \pi_\theta > c \cdot \pi_{\text{old}} \quad \text{(Dual Clip)} \end{cases} \tag{A3}$$

This derivation explains the blue regions in Fig. 1(c). The Cap is maximized when π_θ is near the dual-clip boundary and gradually diminishes to zero as π_θ approaches the trust-region lower limit.

C Derivation of the UP-GSPO Gradient

To rigorously establish the gradient behavior of UP-GSPO for positive advantage samples ($\hat{A}_i > 0$), we derive the exact analytical gradient of its sequence-level surrogate objective.

For a given prompt q and a set of G sampled responses $\{o_i\}_{i=1}^{G}$, we define the self-anchored sequence-level Unbounded Positive (UP) objective as follows:

$$\mathcal{J}^{+}_{\text{UP-GSPO}}(\theta) = \mathbb{E}_{q\sim\mathcal{Q},\{o_i\}_{i=1}^{G}\sim\pi_{\text{old}}}\left[\frac{1}{G}\sum_{i=1}^{G}\hat{A}_i\left(\frac{\pi_\theta(o_i|q)}{\text{sg}(\pi_\theta(o_i|q))}\right)^{\frac{1}{|o_i|}}\right] \tag{A4}$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operator, which strictly treats its operand as a constant during backpropagation.

When applying the gradient operator ∇_θ, the denominator $\text{sg}(\pi_\theta(o_i|q))$ factors out as a constant. Because the forward numerical value of the stop-gradient term is strictly identical to the policy probability (i.e., $\text{sg}(\pi_\theta) \equiv \pi_\theta$), we can seamlessly substitute this equivalence back into the expression after applying the chain rule $\nabla_\theta(f^\alpha) = \alpha f^{\alpha-1}\nabla_\theta f$. This yields a continuous and elegant derivation:

$$\begin{aligned}
\nabla_\theta\mathcal{J}^{+}_{\text{UP-GSPO}}(\theta) &= \mathbb{E}_{q,\{o_i\}}\left[\frac{1}{G}\sum_{i=1}^{G}\hat{A}_i\frac{1}{\text{sg}(\pi_\theta(o_i|q))^{\frac{1}{|o_i|}}}\nabla_\theta\left(\pi_\theta(o_i|q)^{\frac{1}{|o_i|}}\right)\right] \\
&= \mathbb{E}_{q,\{o_i\}}\left[\frac{1}{G}\sum_{i=1}^{G}\hat{A}_i\frac{1}{\pi_\theta(o_i|q)^{\frac{1}{|o_i|}}}\left(\frac{1}{|o_i|}\pi_\theta(o_i|q)^{\frac{1}{|o_i|}-1}\nabla_\theta\pi_\theta(o_i|q)\right)\right] \\
&= \mathbb{E}_{q,\{o_i\}}\left[\frac{1}{G}\sum_{i=1}^{G}\hat{A}_i\frac{1}{|o_i|}\frac{\nabla_\theta\pi_\theta(o_i|q)}{\pi_\theta(o_i|q)}\right]
\end{aligned} \tag{A5}$$

Notice how the complex length-normalization exponents algebraically cancel out perfectly ($-\frac{1}{|o_i|}+\frac{1}{|o_i|}-1=-1$).

Finally, by applying the log-derivative trick ($\frac{\nabla_\theta\pi}{\pi} = \nabla_\theta\log\pi$) and expanding the sequence-level joint probability into the sum of its token-level log probabilities ($\log\pi_\theta(o_i|q) = \sum_{t=1}^{|o_i|}\log\pi_\theta(o_{i,t}|q,o_{i,<t})$), the expected gradient simplifies to our final form:

$$\nabla_\theta\mathcal{J}^{+}_{\text{UP-GSPO}}(\theta) = \mathbb{E}_{q\sim\mathcal{Q},\{o_i\}_{i=1}^{G}\sim\pi_{\text{old}}}\left[\frac{1}{G}\sum_{i=1}^{G}\hat{A}_i\left(\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\nabla_\theta\log\pi_\theta(o_{i,t}|q,o_{i,<t})\right)\right] \tag{A6}$$

Remark: Equation A6 rigorously demonstrates that for positive advantages, the UP-GSPO sequence-level objective mathematically equates to a length-normalized REINFORCE gradient. By completely bypassing the dynamic importance sampling ratio, this formulation explicitly removes the clipping upper bound, maximizing exploration capacity while cleanly preserving the variance-reducing properties of the $\frac{1}{|o_i|}$ normalization.

Building upon this derivation, we formulate the final, unified UP-GSPO objective. Recognizing the divergent optimization dynamics between correct and incorrect rollouts, UP-GSPO employs an asymmetric routing mechanism. We retain the standard sequence-level clipped objective for negative advantages to act as a structural safeguard, while deploying our Unbounded Positive objective for positive advantages.

Let $s_i(\theta) = \left(\frac{\pi_\theta(o_i|q)}{\pi_{\text{old}}(o_i|q)}\right)^{\frac{1}{|o_i|}}$ denote the length-normalized sequence-level importance weight. The final UP-GSPO objective is formulated as:

$$\mathcal{J}_{\text{UP-GSPO}}(\theta) = \mathbb{E}_{q\sim\mathcal{Q},\{o_i\}\sim\pi_{\text{old}}}\left[\frac{1}{G}\sum_{i=1}^{G}\begin{cases}\hat{A}_i\left(\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\log\pi_\theta(o_{i,t}|q,o_{i,<t})\right) & \text{if } \hat{A}_i > 0 \\ \min\left(s_i(\theta)\hat{A}_i,\ \text{clip}(s_i(\theta),1-\epsilon,1+\epsilon)\hat{A}_i\right) & \text{if } \hat{A}_i \le 0\end{cases}\right] \tag{A7}$$