WHEN IMPORTANCE SAMPLING MISALLOCATES CREDIT: ASYMMETRIC RATIOS FOR OUTCOME-SUPERVISED RL

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning (RL) has shown great promise in large language models (LLMs) post-training, which typically rely on token-level clipping to maintain stability during optimization. Despite the empirical success of GRPO-style methods, we identify a fundamental and previously overlooked challenge in this popular Outcome-Supervised RL (OSRL) paradigm. We reveal that in OSRL, where advantages are shared across tokens within a response, **importance sampling (IS) ratios also act as token-level weights beyond their traditional role of distribution correction** that allocate the shared advantage signal across tokens. We show that this hidden role shift induces a critical mismatch for positive-advantage tokens, leading to unbalanced token weighting between positive and negative tokens. Specifically, it suppresses the update of underrepresented tokens that are lagging behind, while over-amplifying already high-probability tokens. This mismatch results in rich-get-richer dynamics that over-reinforce confident tokens, weaken catch-up learning that drive entropy collapse, excessive repetition, and premature convergence. To address this, we propose **A**symmetric Importance **S**ampling **P**olicy **O**ptimization (ASPO), a simple yet effective strategy that reverses the ratio-induced weighting of positive-advantage tokens, while stabilizing extreme updates and maintaining gradient flow for unmasked tokens. This mismatch correction aligns their weighting trend with the learning dynamics of negative ones. Comprehensive experiments across math reasoning and coding benchmarks demonstrate that ASPO significantly mitigates entropy collapse, improves training stability, and enhances performance over strong GRPO-based baselines. Our analysis provides new insights into the role of token-level weighting in OSRL and highlights the critical importance of correcting ratio-induced weighting in LLM RL.

1 INTRODUCTION

Reinforcement Learning (RL) has achieved remarkable success across various domains, including board games (Silver et al., 2017; 2018), robotic control (Kober et al., 2013; Andrychowicz et al., 2020; Chen et al., 2022), Large Language Model (LLM) alignment (Ouyang et al., 2022; Bai et al., 2022) and reasoning (Guo et al., 2025; OpenAI, 2024; Yu et al., 2025). In classical RL, an agent interacts with an environment and receives intermediate rewards at different timesteps along a trajectory, where different actions contribute unequally to the final return, leading to fine-grained credit assignment for effective policy learning.

In this regime, policy optimization methods such as Proximal Policy Optimization (PPO) (Schulman et al., 2017) rely on importance sampling (IS) ratios to account for the distribution mismatch between the behavior policy that generated the data and the current policy being optimized. Together with the clipping mechanism, IS plays a principled role in reusing on-policy samples for optimization while preventing excessively large policy changes. In standard PPO, the IS ratio is coupled with a per-step advantage A_t, and each action-specific ratio modulates the contribution of the corresponding action-specific advantage (Figure 1(a)). Thus, the probability ratio primarily serves as a distribution-correction factor within a fine-grained credit-assignment structure.

Outcome-Supervised Reinforcement Learning (OSRL) for LLMs departs sharply from the traditional RL paradigm (Zhang et al., 2025). Group Relative Policy Optimization (GRPO) (Shao et al., 2024) and its variants, such as DAPO (Yu et al., 2025), have become widely adopted in this setting due

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.

to their simplicity and effectiveness. They estimate a response-level advantage from group rollouts $\hat{A}_{\text{resp}}$ which is *shared* across all tokens in the response. Despite its empirical success, GRPO inherits optimization machinery from PPO (e.g., IS ratios) in a regime that is substantially different from the one for which PPO was originally designed. We uncover a previously underexplored consequence of this design: under the commonly-adopted shared advantage setups with $\hat{A}_{\text{resp}}$ in OSRL, the distribution-correction interpretation alone does not explain the token-weighting effect of PPO-style ratios. Since $\hat{A}_{\text{resp}}$ is shared across all tokens, it no longer provides token-level specific credit assignment, and the ratio also acts as intra-response ***token-level weights*** that determine how strongly each token contributes to the gradient update.

This role shift leads to a critical mismatch in token weighting, as shown in Figure 4. When the old-policy probability is fixed, for positive-advantage samples, tokens that already have higher probability under the current policy receive larger update weights, while tokens with low probability are heavily down-weighted. This creates a richer-get-rich dynamic: already confident tokens in successful responses are reinforced more aggressively, while lagging but desirable tokens are under-updated when they may need stronger learning signals to catch up, as illustrated in Figure 1(b), which has been overlooked in previous research. Existing modifications to GRPO, such as clipping-threshold adjustments in DAPO, partially relax the upper clipping bound to facilitate low-probability token learning, but does not address the core issue that low-probability positive tokens remain weakly updated. This design contradicts the desired learning dynamics, where low-probability tokens in positive trajectories should receive stronger updates to catch up, and result in a series of abnormal training dynamics. Empirically, we observe accelerated entropy collapse, increasing repetition rate, rising clipping ratios, and premature convergence (Figure 2). These phenomena indicate unstable optimization driven by self-reinforcing updates on confident tokens, rather than healthy convergence.

To address these issues, we propose **A**symmetric Importance **S**ampling **P**olicy **O**ptimization (ASPO), a simple yet effective method for GRPO-style OSRL. ASPO is based on the observation that, once a shared response-level advantage is broadcast to all tokens, the ratio primarily determines how this outcome signal is distributed across tokens. For positive-advantage responses, ASPO reverses the direction of ratio-induced token weighting: low-probability positive tokens receive stronger updates, while already confident tokens are down-weighted. This directly counteracts the rich-get-richer effect of standard GRPO, as shown in Figure 1. To further stabilize training, we incorporate a soft dual-clipping mechanism (Ye et al., 2020; Chen et al., 2025) that constrains extreme ratios without discarding gradients for positive tokens. Extensive experiments on both mathematical reasoning and coding benchmarks demonstrate that ASPO: (1) mitigates entropy collapse and overfitting, (2) yields more stable training dynamics, and (3) significantly outperforms GRPO-based baselines in final performance.

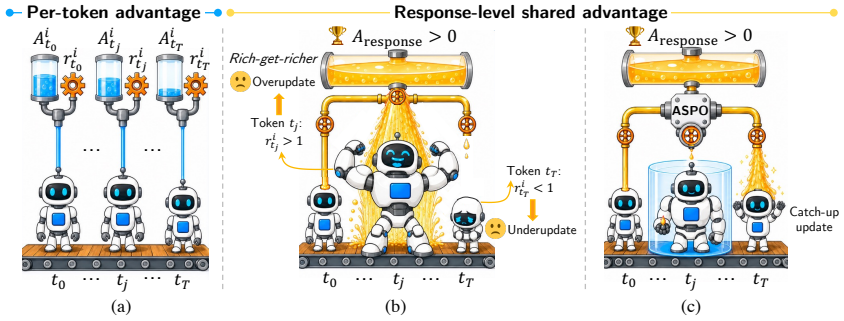


Figure 1: Illustration of ratio-induced token misallocation in outcome-supervised RL and the ASPO correction. For positive tokens, ASPO prevents overoptimization of high-probability tokens and facilitates low-probability token learning.

In summary, our main contributions are as follows:

- We identify a role shift of IS ratios in widely-adopted OSRL and a fundamental misallocation, where the IS ratio acts as token-level weighting and can misallocate credit across tokens: *positive-advantage tokens are weighted in the undesired direction*, with confident tokens overoptimized and lagging tokens suppressed.
- We propose ASPO, an asymmetric ratio-based mechanism that promotes catch-up learning for underrepresented successful tokens while preserving gradient flow.

- We provide extensive empirical evidence that correcting token-level weighting significantly improves training stability and performance across multiple math reasoning and coding benchmarks.

2 Related Work

RL for Large Language Models. Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017; Liu et al., 2022) has achieved remarkable success in aligning LLMs with human values. Recently, DeepSeek-R1 (Guo et al., 2025) and the GRPO algorithm (Shao et al., 2024) have demonstrated that RLVR effectively enhances the reasoning capabilities of LLMs. Subsequent works have extended the GRPO algorithm for RLVR (Luo et al., 2025b; Yu et al., 2025; Yue et al., 2025; He et al., 2025; Wang et al., 2025; Liu et al., 2025b; Su et al., 2025a;b; Zhu et al., 2026). Our method follows this GRPO-based line of research but introduces a new approach of the clipping and IS mechanism to address inherent limitations in GRPO.

Clipping Mechanism in RL. PPO (Schulman et al., 2017) introduces clipping based on importance sampling ratios as an alternative to KL divergence to constrain policy updates relative to the reference policy, and GRPO (Shao et al., 2024) adopts this clipping loss for LLM RL. There has been a growing line of work investigating whether this clipping is the right stabilization mechanism for LLM RL. CISPO (Chen et al., 2025) observes that gradients for clipped tokens are masked and proposes preserving them based on the PPO-Clip objective (Schulman et al., 2017) but without the conservative $\min$ operation. GSPO (Zheng et al., 2025) argues that the optimization objective should match the sequence-level reward's granularity, proposing sequence-level clipping and IS ratio computation. DCPO (Yang et al., 2025b) employs dynamic-adaptive clipping ranges instead of fixed bounds. Our method also targets the clipping term in GRPO but differs primarily by focusing on the IS ratio, incorporating a reciprocal weight for positive tokens. DAPO (Yu et al., 2025) and DPPO (Qi et al., 2026) further relax the clipping ranges of low-probability (positive) tokens, but do not address the problem of weight mismatch of positive tokens. These studies mainly refine the gating behavior of ratio-based objectives, deciding when updates should be clipped, masked, relaxed, or computed at a different granularity to maintain training stability. In contrast, our work focuses on the overlooked phenomenon of relative weighting of updates that remain active.

3 The Role Shift of Importance Sampling Ratio in OSRL

3.1 Theoretical Understanding: PPO-style IS Ratios as Induced Token Weights in OSRL

Importance Sampling (IS) enters policy optimization as a way to evaluate an objective under one policy using samples collected from another (Precup et al., 2000) for optimizing $\sum_a \pi_\theta(a|s) A_{\theta_{\text{old}}}(s,a)$, where θ and θ_{old} denote parameters of the current and old policies, respectively, with A the stepwise advantage function for the current step's state action pair (s,a). Since trajectories are sampled from the old policy $\pi_{\theta_{\text{old}}}$, this expectation is rewritten using an importance sampling estimator $\mathbb{E}_{a\sim\pi_{\theta_{\text{old}}}(\cdot|s)}\left[\frac{\pi_\theta(a|s)}{\pi_{\theta_{\text{old}}}(a|s)} A_{\theta_{\text{old}}}(s,a)\right]$. This derivation gives the likelihood ratio $r_t = \frac{\pi_\theta}{\pi_{\text{old}}}$ a clear interpretation, which corrects the distribution mismatch between the old policy that generated the data and the current policy being optimized for that action (Schulman et al., 2017). In the classic PPO surrogate in the RL literature, this IS ratio is coupled with a step-specific advantage A_t, which specifies the estimated credit of the same timestep t (operating at the same granularity).

In OSRL for LLMs, such as GRPO (Shao et al., 2024) and its variants (e.g., DAPO (Yu et al., 2025)), this granularity alignment is broken. GRPO-style methods still use token-level probability ratios r_t, but the advantage is not estimated at the token level. Instead, a response-level advantage $\hat{A}_{\text{resp}}$ is estimated and then broadcast to every token in the response, i.e., $A_t^i = \hat{A}^i, \forall t$. Thus, all tokens within a response *share the same advantage value*, which is fundamentally different from the classical setting where A_t provides timestep-specific credit. Once the same $\hat{A}_{\text{resp}}$ is assigned to every token, the advantage can no longer distinguish which tokens should receive stronger or weaker updates.

The original motivation of IS is to correct the distribution mismatch of the action-specific advantage term. However, under shared advantages, there is no token-specific credit estimation, which makes it unclear what the token-level ratio is correcting when the advantage itself no longer provides token-level credit, leading to the following critical question:

If the real credit of each token is already unclear due to outcome-based advantage estimation, what is the role and effect of further adjusting the distribution using IS weights?

We first answer this question theoretically by analyzing the gradient of GRPO, which is further empirically validated in Section 3.2. The following proposition shows that it is gradient-equivalent to a weighted log-likelihood objective with ratio-induced token weights.

Proposition 3.1 (IS ratio as token-wise gradient weighting in GRPO)**.** *Consider the GRPO objective in Eq. (3). For each token o_t^i, define the active clipping mask* $m_t^i = \begin{cases} 1, & A_t^i \geq 0 \text{ and } r_t^i \leq 1+\epsilon, \\ 1, & A_t^i < 0 \text{ and } r_t^i \geq 1-\epsilon, \\ 0, & \text{otherwise} \end{cases}$ *and the detached token weight $w_t^i = \mathrm{sg}(m_t^i r_t^i)$, where $\mathrm{sg}(\cdot)$ denotes the stop-gradient operator. Consider the weighted log-likelihood objective in Eq. (1), then we have $\nabla_\theta \mathcal{J}_{\mathrm{GRPO}}(\theta) = \nabla_\theta \mathcal{J}_{\mathrm{WLL}}(\theta)$.*

$$\mathcal{J}_{WLL}(\theta) = \mathbb{E}_{q\sim\mathcal{D},\{o^i\}_{i=1}^{G}\sim\pi_{\theta_{old}}(\cdot|q)} \left[\frac{1}{G}\sum_{i=1}^{G} \frac{1}{|o^i|} \sum_{t=1}^{|o^i|} \left(w_t^i A_t^i \log \pi_\theta(o_t^i|q, o_{<t}^i) - \beta \mathbb{D}_{KL}(\pi_\theta \| \pi_{ref}) \right) \right]. \tag{1}$$

Remark. Proposition 3.1 reveals how the IS ratio enters the GRPO update, and the proof can be found in Appendix B.1. Because GRPO computes the advantage at the response level and broadcasts it to all tokens, the token-level ratio r_t^i is no longer paired with a token-specific credit estimate. Instead, after clipping, it becomes the main token-dependent scalar that modulates the update of each token on the shared advantage, i.e., $\mathrm{sg}(m_t^i r_t^i)\hat{A}^i$. Therefore, tokens within the same response can receive different scalar update weights solely because their current-to-old probability ratios differ, rather than because their individual contributions to the response reward are different.

3.2 Experimental Validation

The gradient view in Appendix B.2 motivates our empirical analysis. Although IS ratios are introduced to compensate for the mismatch between the behavior policy and the current policy, GRPO-style OSRL computes advantages at the response level and broadcasts them to all tokens. Thus, the IS ratio becomes the main token-specific factor in the effective update weight. We therefore ask whether IS is necessary as a distribution-correction mechanism in this setting, or whether it mainly reshapes learning dynamics through ratio-induced weighting.

Setup. To evaluate the practical effects of IS weights in OSRL, we compare two variants that are identical except for the use of r_t ratios: (1) GRPO with original IS weights, and (2) GRPO without IS weights (where the detached ratio factor in the weighted log-likelihood formulation is fixed to 1.0). If IS were indispensable for correcting this mismatch, removing it would be expected to noticeably degrade performance or destabilize training.

3.2.1 Results

Figure 2 compares the evaluation accuracy and training dynamics of standard GRPO and GRPO without IS. The two variants achieve comparable final test accuracy. Standard GRPO reaches its peak slightly earlier and stops to increase, whereas GRPO without IS converges more smoothly with competitive performance. Its training dynamics are also more stable, with slower entropy decay and smaller increases in repetition, and KL divergence. These results suggest that removing IS does not compromise final performance, but makes optimization less aggressive.

3.2.2 Analysis

To understand what the IS ratios are doing, we further examine their values during GRPO training. Figure 3 shows the response-level IS ratios, where the average IS ratio is typically slightly greater than 1.0 (around 1.0004), and the ratios are not uniformly distributed across samples. Specifically, the average IS ratio of positive samples is slightly higher than that of negative samples. While the numerical gap (for each update) is small, it accumulates over many tokens and repeated off-policy updates, and its cumulative effect is significant. According to Proposition 3.1, the GRPO gradient can be written as a weighted log-likelihood update whose effective weight is proportional to $r_{i,t}\hat{A}_i$, which changes the relative strength of the corresponding gradient terms. For positive responses, ratios above one amplify the reward-increasing update. For negative responses, ratios below one tend to shrink the magnitude of the penalizing update in the gradient-active region. Thus, it causes GRPO

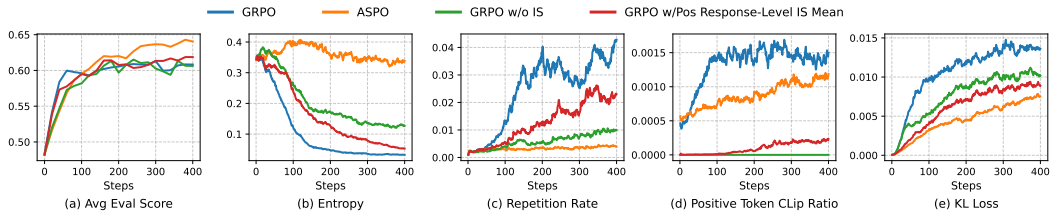


Figure 2: The training dynamics curves of all methods on Qwen3-4B, including (a) average evaluation accuracy, (b) entropy, (c) repetition rate, (d) positive token clip ratio, and (e) KL loss. The curves are smoothed with EMA for better visualization.

training to **prioritize learning positive responses rather than suppressing negative ones**, and this gap widens across training, which amplifies positive-sample learning and accelerates entropy decay.

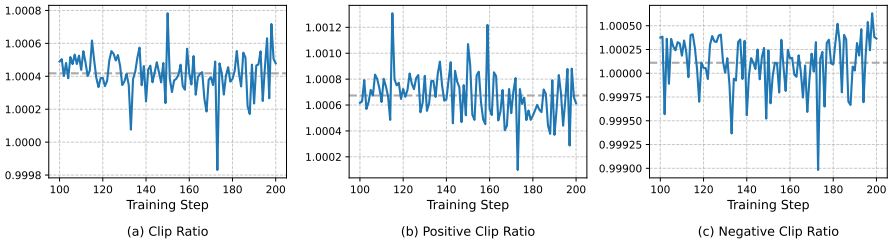


Figure 3: Curves of response-level IS ratios throughout GRPO training. The average IS ratios are shown in gray dashed lines.

What happens without IS weights? With all IS ratios set to 1.0, the update no longer contains this ratio-induced preference amplification. On the one hand, this makes the overall weights slightly lower than in standard GRPO. On the other hand, it *eliminates the ratio-induced weight difference between positive and negative samples*. While these two factors slow down the learning speed compared to standard GRPO, they do not compromise the final performance, as shown in Figure 2.

This observation supports the role-shift interpretation in Section 3.1. In GRPO-style OSRL, although IS ratios still quantify the deviation between the current policy and the behavior policy, when a response-level advantage is shared by all tokens, their dominant operational effect is to reallocate the outcome-level learning signal across samples and tokens. The more aggressive entropy decay, repetition growth, clipping, and KL increase observed in Figure 2 are consistent with this ratio-induced amplification effect, and motivate the finer-grained token-level analysis in the next section.

Based on the above analysis, we draw the following conclusions:

Takeaways for Importance Sampling

- In the shared-advantage GRPO setting, removing ratio-induced weights preserves comparable final performance, while making optimization smoother and less aggressive.
- Although IS ratios still measure current-to-old policy deviation, their dominant operational role is to act as effective update weights under shared response-level advantages.
- IS ratios induce an asymmetric weighting pattern: positive responses are amplified more than negative responses are suppressed.

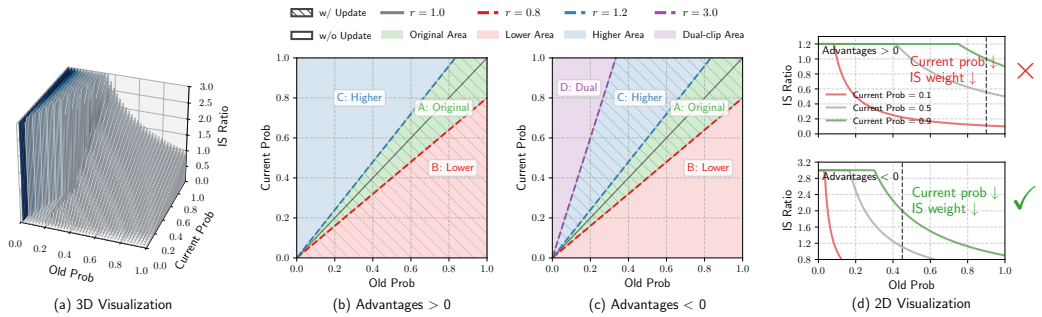
4 Ratio-induced Weight Mismatch in Positive-Advantage Responses

4.1 Rethinking the Role of IS-induced Token Weights

The analysis in Section 3 shows that, under shared response-level advantages, PPO-style IS ratios (after clipping) in GRPO-style OSRL primarily act as **token-level training weights**, which modulate

how strongly each token is updated under the same response-level advantage. This motivates a different question: *if the ratio is viewed as an update weight, does the PPO-Clip weighting pattern allocate learning signal to tokens in a desirable way?*

The design principle of PPO-Clip is to ensure training stability by preventing tokens that already have a strong advantage in the update direction from dominating the update. This avoids overly aggressive parameter changes that could push the model too far from the old policy. For positive-advantage tokens, an ideal weighting scheme might look like this: the lower a token's probability is relative to the old policy, the larger its training weight should be. Conversely, the higher its probability, the smaller the weight should be. There are two reasons for this: (1) Assigning higher weight to tokens that are lagging behind accelerates their learning progress. (2) For tokens whose probabilities have fallen below the old-policy values, increasing their probabilities initially moves them back toward those values. To make this more intuitive, we visualize IS weights in a three-dimensional coordinate plot, where the z-axis represents the original IS ratio, as shown in Figure 4(a). It can be seen that when the current probability is low, the corresponding IS weight is also small, resulting in insufficient training for these tokens.



(a) 3D Visualization (b) Advantages > 0 (c) Advantages < 0 (d) 2D Visualization

Figure 4: Visualization of IS weights in PPO-Clip. (a) 3D visualization of IS weights, showing that for a fixed old probability, tokens with lower current probabilities receive smaller IS weights. (b) When the advantage is positive ($A > 0$), GRPO optimizes regions A and B. (c) When the advantage is negative ($A < 0$), GRPO optimizes regions A and C. (d) 2D visualization of IS weights, illustrating that along a fixed old probability (dashed line), IS weights monotonically decrease as the current probability decreases. For negative-advantage tokens, this behavior aligns with our expectations, whereas for positive-advantage tokens, it contradicts the desired optimization trend.

4.2 Weight Misallocation of Positive Tokens in PPO-Clip

As shown in Figure 4(c), for negative-advantage samples, the weight distribution behaves as expected: weights decrease gradually from the top-left region to the bottom-right region. However, for positive-advantage samples shown in Figure 4(b), the allocation is the opposite to our intuition. Tokens in the top-left region, whose probabilities under the current policy are already much larger than under the old policy, are given larger weights, while tokens in the bottom-right region, with lower current probabilities, are assigned very small weights. Therefore, PPO-Clip underweights lagging tokens and overweights already reinforced tokens within positive-advantage samples.

This mismatch causes two problems: (1) tokens in the bottom-right region, which are clearly lagging behind the old policy, are suppressed further by excessively low weights. For example, if the old policy probability is 0.9 and the current policy probability is 0.1, the assigned weight is merely $1/9$, resulting in a weak update signal and insufficient learning. (2) for unmasked tokens with $1 < r_t^i < 1 + \varepsilon$ that have already been reinforced, they are assigned disproportionately high weights. On the one hand, the model is more likely to deviate from the old policy, undermining training stability. On the other hand, the model overfits on positive samples, further amplifying their probabilities after the update. In subsequent updates on the same rollout batch, their weights can increase until clipping activates, forming a **self-reinforcing loop**. This mechanism underlies the previously observed phenomena such as entropy collapse and increased output repetition. For more details on how to distinguish healthy convergence and local optima, please refer to Appendix E.

4.3 Experimental Validation

Setup. To empirically validate the preceding analysis, we conduct a controlled comparison between the original GRPO baseline and a modified variant. In this variant, we replace the token-level IS ratios of **positive-advantage samples** with their **response-level average IS ratios**, while keeping the negative-advantage samples unchanged. This design isolates the impact of the mismatched IS weights identified in Section 4.2, ensuring that any observed behavioral differences arise solely from the reweighting of positive tokens. If our hypothesis is correct, the modified setup should partially alleviate the weight mismatch issue.

4.3.1 Results

Improved exploration and smoother training dynamics. After replacing the positive-token IS weights with response-level means, all training curves become substantially smoother. The entropy decline slows down (Figure 2), preventing premature convergence to local optima, and encouraging more diverse and exploratory behavior of the policy. Additionally, the increases in repetition rate, positive clipping ratio, and KL loss are noticeably moderated without the accelerating trends observed in GRPO. These results provide strong empirical evidence that the original IS ratio design indeed leads to unstable optimization through excessive weighting of high-probability tokens.

Better performance with stable training. As shown in Figure 2, the modified method achieves performance improvements compared with GRPO, demonstrating that stable training does not compromise final performance.

Takeaways for IS Weights of Positive Tokens

- The standard PPO-Clip design introduces a **token-weight mismatch for positive-advantage samples**, where high-probability tokens receive disproportionately large update weights.
- This imbalance leads to entropy collapse, and increased repetition, ultimately pushing the policy toward a local optimum and limiting its capacity for continual improvement.
- Reweighting positive samples with response-level IS ratios effectively mitigates these issues, confirming the validity of our preliminary analysis and motivating the design of our proposed method in the next section.

5 Asymmetric Importance Sampling Policy Optimization

Based on the analysis in Section 3 and 4, we show that token-level IS ratios in GRPO-style OSRL should be understood primarily as token-wise update weights under shared response-level advantages. In addition, the response-level mean-ratio diagnostic in Section 5 shows that removing token-level ratio variation from positive-advantage responses substantially stabilizes training. This perspective reveals an asymmetric issue: the original PPO-style ratio weighting is reasonably aligned with negative-advantage responses, but mismatched for positive-advantage responses. For $\hat{A} > 0$, tokens with $\pi_\theta \ll \pi_{\text{old}}$ are lagging behind the desired update direction but receive small weights, while tokens with $\pi_\theta \gg \pi_{\text{old}}$ have already been reinforced but receive large weights.

Motivated by this, we propose **A**symmetric Importance **S**ampling **P**olicy **O**ptimization (ASPO), a simple yet effective approach for clipping and IS ratio computation, following the mismatch analysis in Section 4. ASPO inverts the IS weights of positive samples, aligning their update behavior with that of negative samples. In other words, tokens whose current policy probabilities are lower than the old policy should be assigned higher learning weights, while those with higher probabilities should receive lower weights.

Specifically, the implementation can be divided into three steps:

Step 1: Token Masking. We retain the original clipping mechanism in GRPO. The policy-gradient contribution (excluding the KL term) of tokens that satisfy: (1) $r_t^i(\theta) < 1 - \varepsilon_{\text{low}}$ ($\hat{A}_t^i < 0$) or (2) $r_t^i(\theta) > 1 + \varepsilon_{\text{high}}$ ($\hat{A}_t^i > 0$) will be masked in a hard clipping manner.

Step 2: Weight Flipping. For tokens with negative advantage values, the ASPO ratio is the same as that of GRPO, i.e., $\hat{r}_t^i = r_t^i$. For tokens with $\hat{A}_t^i > 0$, we use the reciprocal of their IS weights and

Table 1: Evaluation results on mathematical benchmarks. The results of ASPO are shaded and the highest values are **bolded**.

Method	AIME24		AIME25		AMC23		MATH-500		Minerva		Olympiad		Avg.
	avg@64	pass@64	avg@64	pass@64	avg@64	pass@64	avg@4	pass@4	avg@8	pass@8	avg@4	pass@4	
DeepSeek-R1-1.5B	30.6	80.0	23.5	63.3	70.7	**100.0**	83.6	92.4	27.6	48.2	44.6	59.4	46.8
↳ GRPO	42.1	80.0	28.6	56.7	80.3	97.5	87.6	**94.6**	29.2	46.3	53.2	65.8	53.5
↳ DeepScaleR-1.5B	42.0	**83.3**	29.0	63.3	81.3	**100.0**	87.7	93.6	30.3	**51.1**	50.7	61.0	53.5
↳ Nemotron-1.5B	48.0	76.7	33.1	60.0	86.1	97.5	**90.6**	93.6	**35.3**	47.8	**59.2**	66.8	58.7
↳ ASPO-Math-1.5B	**49.0**	80.0	**35.1**	**70.0**	**87.2**	95.0	90.5	94.4	35.1	50.4	58.8	**66.9**	**59.3**
Qwen3-4B	23.6	56.7	18.3	63.3	67.7	95.0	84.5	92.4	41.5	56.3	54.1	66.6	48.3
↳ GRPO	43.4	**83.3**	35.5	**70.0**	84.3	**97.5**	91.7	95.8	47.2	58.5	67.4	75.8	61.6
↳ ASPO-Math-4B	**50.5**	**83.3**	**40.9**	**70.0**	**87.4**	**97.5**	**93.4**	**97.0**	**49.0**	**60.7**	**68.8**	**78.5**	**65.0**
Qwen3-8B	27.0	63.3	19.1	56.7	68.9	**97.5**	83.6	92.4	43.9	58.1	55.7	69.6	49.7
↳ GRPO	50.3	**83.3**	34.1	66.7	84.1	95.0	92.7	96.0	50.2	**61.4**	68.2	76.1	63.3
↳ ASPO-Math-8B	**52.4**	**83.3**	**38.9**	**80.0**	**89.1**	**97.5**	**93.6**	**96.4**	**50.6**	60.7	**69.8**	**77.9**	**65.7**
Qwen3-30B-A3B	30.1	70.0	19.9	56.7	74.3	**100.0**	88.4	96.0	47.7	59.6	59.6	71.2	53.3
↳ GRPO	59.5	80.0	43.1	73.3	91.9	97.5	95.3	97.8	51.9	61.8	71.0	79.2	68.8
↳ ASPO-Math-30B	**61.7**	**90.0**	**50.3**	**80.0**	**94.9**	97.5	**95.7**	**98.4**	**54.3**	**64.3**	**74.3**	**81.5**	**71.8**

the ASPO ratio is computed as:

$$\hat{r}_t^i = \frac{\pi_{\theta_{\text{old}}}(o_t^i \mid q, o_{<t}^i)\pi_\theta(o_t^i \mid q, o_{<t}^i)}{\text{sg}(\pi_\theta^2(o_t^i \mid q, o_{<t}^i))}, \tag{2}$$

where $\text{sg}(\cdot)$ denotes stop gradient operation. The gradient analysis in Appendix B.2 shows that the gradient of positive tokens in ASPO is positively correlated with $\frac{1}{\pi_\theta}$, indicating that the gradient becomes larger when the probability is lower.

Step 3: Dual Clipping. PPO-Clip can be extended with a dual-clip mechanism (Ye et al., 2020) to handle cases where, for $\hat{A} < 0$, extremely large ratios could lead to weight explosion, destabilizing training. Originally, for the $\hat{A} > 0$ region, this problem was naturally avoided by the hard clipping mechanism. However, since we now invert the weights for positive samples, extreme cases shift to the right-hand side of the $\hat{A} > 0$ region (region B in Figure 4(b)). Therefore, when using ASPO, positive-sample tokens also require dual-clip. Specifically, this dual-clip is implemented using a **soft clipping** manner, which only clips the values but retains the gradient.

It is important to note that tokens clipped by dual-clip are fundamentally different from tokens masked in the first step. The latter are blocked because they already have sufficient advantage in the update direction, whereas the former are tokens that lag significantly behind the old policy but need their weight magnitude constrained due to abnormal computation. We still want these tokens to participate in training, so we use the soft clipping in CISPO (Chen et al., 2025) for these dual-clipped tokens.

6 Experiments

6.1 Setup

Models and Baselines. We conduct experiments using DeepSeek-R1-Distill-Qwen-1.5B (Guo et al., 2025), Qwen3-4B, Qwen3-8B, and Qwen3-30B-A3B[1] (Yang et al., 2025a) as the base model. We compare ASPO against several representative baselines: (1) Base Model, (2) GRPO (Shao et al., 2024), (3) DeepScaleR-1.5B (Luo et al., 2025b), (4) DeepCoder-1.5B (Luo et al., 2025a), and (5) Nemotron-1.5B (Liu et al., 2025a).

Evaluation. We evaluate models on both mathematical and coding domains. For math, we use six challenging datasets: AIME24 (MAA, 2024), AIME25 (MAA, 2025), AMC23 (MAA, 2023), MATH-500 (Lightman et al., 2024), Minerva Math (Lewkowycz et al., 2022), and OlympiadBench (He et al., 2024). For coding, we adopt LiveCodeBench v5 (2024.08.01–2025.02.01) and v6 (2025.02.01–2025.05.01) (Jain et al., 2025). vLLM (Kwon et al., 2023) is used for inference with a maximum output length of 32,768 tokens and a temperature of 0.8. We report both avg@K and pass@K for each benchmark.

[1]We use non-thinking mode for Qwen3 models.

Table 2: Evaluation results on code benchmarks. The results of ASPO are shaded and the highest values are **bolded**.

Method	LCB v5 (2024.08.01-2025.02.01)		LCB v6 (2025.02.01-2025.05.01)		Avg.
	avg@8	pass@8	avg@16	pass@16	
DeepSeek-R1-1.5B	16.7	29.0	17.2	34.4	17.0
↳ GRPO	26.0	40.5	27.6	43.5	26.8
↳ DeepCoder-1.5B	23.3	39.1	22.6	42.0	23.0
↳ Nemotron-1.5B	26.1	35.5	29.5	42.8	27.8
↳ ASPO-Code-1.5B	**31.5**	**47.0**	**30.5**	**46.0**	**31.0**
Qwen3-4B	23.8	35.8	24.0	35.1	23.9
↳ GRPO	40.8	55.2	36.6	45.0	38.7
↳ ASPO-Code-4B	**44.8**	**58.1**	**38.3**	**47.3**	**41.6**

Implementation Details. We implement ASPO based on GRPO (Shao et al., 2024) with verl (Sheng et al., 2025). The training batch size is set to 64, with a mini-batch size of 16. The learning rate is 1.0×10^{-6}. Additional implementation details are provided in Appendix C.

6.2 Main Results

The results in Table 1 and Table 2 show that, when using DeepSeek-R1-Distill-Qwen-1.5B as the base model, ASPO improves upon the base model by 12.5 and 14.0 percentage points on math and coding tasks, respectively. Moreover, ASPO consistently outperforms GRPO and several strong OSRL methods averaged across all evaluated benchmarks, demonstrating the effectiveness of ASPO. On larger dense and MoE models, ASPO still exceeds the baselines by a large margin, demonstrating the effectiveness across different model sizes. As shown in Table 3, ASPO still outperforms the baseline method on average across OOD tasks.

6.3 Analysis

As shown in Figure 2 and Table 1, compared with GRPO, ASPO yields trends that align with our earlier findings, but with even stronger improvements. Specifically, entropy decreases more gradually, the repetition rate and positive token clipping ratio grow more slowly, and training is more stable overall, showing characteristics of healthy convergence as discussed in Appendix E. More training dynamics results are shown in Appendix D.

More importantly, since entropy declines more smoothly and remains at a higher level, the model avoids premature collapse and continues learning effectively. As training progresses, performance steadily improves, significantly surpassing the best results achieved by GRPO-based training.

It is also noteworthy that in the early training stages, models trained with ASPO exhibit slightly slower reward improvement (Figure 5(a) and 6(a)) compared to other variants. This occurs because inverting positive-sample weights reduces the overall average IS weight, leading to slower initial fitting of positive samples. However, as training continues, the performance not only catches up but ultimately surpasses the other approaches.

7 Conclusion

In this paper, we identify a fundamental role shift of IS in GRPO-style OSRL: under shared response-level advantages, token-level IS ratios also act as token-wise update weights in addition to their distribution-correction role. We further reveal a weight mismatch for positive-advantage tokens, where already-confident tokens are over-amplified while lagging tokens are under-updated, leading to entropy collapse and unstable training. To address this, we propose ASPO, which flips the IS ratios of positive tokens and incorporates a soft dual-clipping mechanism to stabilize OSRL for LLMs. Experimental results across mathematical and coding domains demonstrate that our method effectively alleviates token-level weight mismatch, mitigates entropy collapse, and improves training stability, leading to superior model performance than the baselines.

Limitations and Discussion. Our investigation primarily targets the importance sampling mechanism within the widely-adopted GRPO-based methods, which have recently emerged as a popular paradigm for efficient LLM post-training. It is a promising direction to extend our investigation to other RL paradigms, e.g., PPO with token-level IS ratios and advantages. We hope our work serves as a foundation for these broader explorations.

AI USE STATEMENT

In this work, we used generative AI tools for assisting with translation. We have not used generative AI tools for generating synthetic data sets, helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, cleaning and reformatting datasets, supporting qualitative and thematic data analysis, or interpreting results. Additionally, we used generative AI tools for editing a research paper to improve readability. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We have provided experimental details in Section 6 and Appendix C. We will also release our code upon acceptance.

REFERENCES

OpenAI: Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafal Jozefowicz, Bob McGrew, Jakub Pachocki, Arthur Petron, Matthias Plappert, Glenn Powell, Alex Ray, et al. Learning dexterous in-hand manipulation. *The International Journal of Robotics Research*, 39(1):3–20, 2020.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

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.

Yuanpei Chen, Tianhao Wu, Shengjie Wang, Xidong Feng, Jiechuan Jiang, Zongqing Lu, Stephen McAleer, Hao Dong, Song-Chun Zhu, and Yaodong Yang. Towards human-level bimanual dexterous manipulation with reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 5150–5163, 2022.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3828–3850, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.211. URL https://aclanthology.org/2024.acl-long.211/.

Jujie He, Jiacai Liu, Chris Yuhao Liu, Rui Yan, Chaojie Wang, Peng Cheng, Xiaoyu Zhang, Fuxiang Zhang, Jiacheng Xu, Wei Shen, et al. Skywork open reasoner 1 technical report. *arXiv preprint arXiv:2505.22312*, 2025.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free

evaluation of large language models for code. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=chfJJYC3iL.

Jens Kober, J Andrew Bagnell, and Jan Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.

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*, SOSP '23, pp. 611–626, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400702297. doi: 10.1145/3600006.3613165. URL https://doi.org/10.1145/3600006.3613165.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving Quantitative Reasoning Problems with Language Models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 3843–3857. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/18abbeef8cfe9203fdf9053c9c4fe191-Paper-Conference.pdf.

Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, Thomas Hubert, Peter Choy, Cyprien de Masson d'Autume, Igor Babuschkin, Xinyun Chen, Po-Sen Huang, Johannes Welbl, Sven Gowal, Alexey Cherepanov, James Molloy, Daniel J. Mankowitz, Esme Sutherland Robson, Pushmeet Kohli, Nando de Freitas, Koray Kavukcuoglu, and Oriol Vinyals. Competition-level code generation with alphacode. *Science*, 378(6624):1092–1097, 2022. doi: 10.1126/science.abq1158. URL https://www.science.org/doi/abs/10.1126/science.abq1158.

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*, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

Mingjie Liu, Shizhe Diao, Ximing Lu, Jian Hu, Xin Dong, Yejin Choi, Jan Kautz, and Yi Dong. Prorl: Prolonged reinforcement learning expands reasoning boundaries in large language models. *arXiv preprint arXiv:2505.24864*, 2025a.

Runze Liu, Fengshuo Bai, Yali Du, and Yaodong Yang. Meta-reward-net: Implicitly differentiable reward learning for preference-based reinforcement learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 22270–22284. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/8be9c134bb193d8bd3827d4df8488228-Paper-Conference.pdf.

Runze Liu, Jiakang Wang, Yuling Shi, Zhihui Xie, Chenxin An, Kaiyan Zhang, Jian Zhao, Xiaodong Gu, Lei Lin, Wenping Hu, Xiu Li, Fuzheng Zhang, Guorui Zhou, and Kun Gai. Attention as a compass: Efficient exploration for process-supervised rl in reasoning models. *arXiv preprint arXiv:2509.26628*, 2025b.

Michael Luo, Sijun Tan, Roy Huang, Ameen Patel, Alpay Ariyak, Qingyang Wu, Xiaoxiang Shi, Rachel Xin, Colin Cai, Maurice Weber, Ce Zhang, Li Erran Li, Raluca Ada Popa, and Ion Stoica. Deepcoder: A fully open-source 14b coder at o3-mini level. https://pretty-radio-b75.notion.site/DeepCoder-A-Fully-Open-Source-14B-Coder-at-O3-mini-Level-1cf81902c14680b3bee5eb349a512a51, 2025a. Notion Blog. Accessed: 2026-04-30.

Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Y. Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Li Erran Li, Raluca Ada Popa, and Ion Stoica. Deepscaler: Surpassing o1-preview with a 1.5b model by scaling rl. https://pretty-radio-b75.notion.site/DeepS

caleR-Surpassing-O1-Preview-with-a-1-5B-Model-by-Scaling-RL-19681902c1468005bed8ca303013a4e2, 2025b. Notion Blog. Accessed: 2026-04-30.

MAA. American mathematics contest 12 (amc 12), November 2023. URL https://artofproblemsolving.com/wiki/index.php/AMC_12_Problems_and_Solutions. Accessed: 2026-04-30.

MAA. American invitational mathematics examination (aime), February 2024. URL https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions. Accessed: 2026-04-30.

MAA. American invitational mathematics examination (aime), February 2025. URL https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions. Accessed: 2026-04-30.

OpenAI. Learning to reason with llms, 2024. URL https://openai.com/index/learning-to-reason-with-llms. Accessed: 2026-04-30.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

Guilherme Penedo, Anton Lozhkov, Hynek Kydlíček, Loubna Ben Allal, Edward Beeching, Agustín Piqueres Lajarín, Quentin Gallouédec, Nathan Habib, Lewis Tunstall, and Leandro von Werra. Codeforces. https://huggingface.co/datasets/open-r1/codeforces, 2025. Accessed: 2026-04-30.

Doina Precup, Richard S. Sutton, and Satinder P. Singh. Eligibility traces for off-policy policy evaluation. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pp. 759–766, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc. ISBN 1558607072.

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.

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, Y 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*, EuroSys '25, pp. 1279–1297, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400711961. doi: 10.1145/3689031.3696075. URL https://doi.org/10.1145/3689031.3696075.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419): 1140–1144, 2018.

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

Zhenpeng Su, Leiyu Pan, Minxuan Lv, Yuntao Li, Wenping Hu, Fuzheng Zhang, Kun Gai, and Guorui Zhou. Ce-gppo: Coordinating entropy via gradient-preserving clipping policy optimization in reinforcement learning. *arXiv preprint arXiv:2509.20712*, 2025b.

Jiakang Wang, Runze Liu, Fuzheng Zhang, Xiu Li, and Guorui Zhou. Stabilizing knowledge, promoting reasoning: Dual-token constraints for rlvr. *arXiv preprint arXiv:2507.15778*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Shihui Yang, Chengfeng Dou, Peidong Guo, Kai Lu, Qiang Ju, Fei Deng, and Rihui Xin. Dcpo: Dynamic clipping policy optimization. *arXiv preprint arXiv:2509.02333*, 2025b.

Deheng Ye, Zhao Liu, Mingfei Sun, Bei Shi, Peilin Zhao, Hao Wu, Hongsheng Yu, Shaojie Yang, Xipeng Wu, Qingwei Guo, Qiaobo Chen, Yinyuting Yin, Hao Zhang, Tengfei Shi, Liang Wang, Qiang Fu, Wei Yang, and Lanxiao Huang. Mastering complex control in moba games with deep reinforcement learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04): 6672–6679, Apr. 2020. doi: 10.1609/aaai.v34i04.6144. URL https://ojs.aaai.org/index.php/AAAI/article/view/6144.

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.

Yu Yue, Yufeng Yuan, Qiying Yu, Xiaochen Zuo, Ruofei Zhu, Wenyuan Xu, Jiaze Chen, Chengyi Wang, TianTian Fan, Zhengyin Du, et al. Vapo: Efficient and reliable reinforcement learning for advanced reasoning tasks. *arXiv preprint arXiv:2504.05118*, 2025.

Kaiyan Zhang, Yuxin Zuo, Bingxiang He, Youbang Sun, Runze Liu, Che Jiang, Yuchen Fan, Kai Tian, Guoli Jia, Pengfei Li, Yu Fu, Xingtai Lv, Yuchen Zhang, Sihang Zeng, Shang Qu, Haozhan Li, Shijie Wang, Yuru Wang, Xinwei Long, Fangfu Liu, Xiang Xu, Jiaze Ma, Xuekai Zhu, Ermo Hua, Yihao Liu, Zonglin Li, Huayu Chen, Xiaoye Qu, Yafu Li, Weize Chen, Zhenzhao Yuan, Junqi Gao, Dong Li, Zhiyuan Ma, Ganqu Cui, Zhiyuan Liu, Biqing Qi, Ning Ding, and Bowen Zhou. A survey of reinforcement learning for large reasoning models. *arXiv preprint arXiv:2509.08827*, 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*, 2025.

Xuekai Zhu, Daixuan Cheng, Dinghuai Zhang, Hengli Li, Kaiyan Zhang, Che Jiang, Youbang Sun, Ermo Hua, Yuxin Zuo, Xingtai Lv, Qizheng Zhang, Lin Chen, Fanghao Shao, Bo Xue, Yunchong Song, Zhenjie Yang, Ganqu Cui, Ning Ding, Jianfeng Gao, Xiaodong Liu, Bowen Zhou, Hongyuan Mei, and Zhouhan Lin. FlowRL: Matching reward distributions for LLM reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=l0bnTKbm9U.

A PRELIMINARIES

A.1 GROUP RELATIVE POLICY OPTIMIZATION

Group Relative Policy Optimization (GRPO) (Shao et al., 2024) samples a group of G rollouts for advantage estimation: $\hat{A}_t^i = \frac{R^i - \text{mean}(\{R^i\}_{i=1}^G)}{\text{std}(\{R^i\}_{i=1}^G)}$. The loss function of GRPO is defined as:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim\mathcal{D},\{o^i\}_{i=1}^G\sim\pi_{\theta_{\text{old}}}(\cdot|q)} \left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o^i|}\sum_{t=1}^{|o^i|}\left(\min\left(r_t^i(\theta)\hat{A}_t^i, \text{clip}\left(r_t^i(\theta), 1-\varepsilon, 1+\varepsilon\right)\hat{A}_t^i\right) - \beta\mathbb{D}_{\text{KL}}(\pi_\theta\|\pi_{\text{ref}})\right)\right], \tag{3}$$

where $r_t^i = \frac{\pi_\theta(o_t^i|q,o_{<t}^i)}{\pi_{\theta_{\text{old}}}(o_t^i|q,o_{<t}^i)}$ is the Importance Sampling (IS) ratio, and β is a weight for the Kullback-Leibler (KL) divergence between the current policy π_θ and the reference policy π_{ref}.

A.2 PPO CLIPPING AND THE IMPROVEMENTS

The clipping mechanism in GRPO (Shao et al., 2024) is first introduced by PPO-Clip (Schulman et al., 2017). It serves as an simple yet effective constraint to prevent large policy updates. However, it is shown in Chen et al. (2025) that this clipping mechanism clips the value and also masks the gradient of the clipped tokens.

B THEORETICAL ANALYSIS

B.1 PROOF OF PROPOSITION 3.1

Proposition 3.1 (IS ratio as token-wise gradient weighting in GRPO)**.** *Consider the GRPO objective in Eq. (3). For each token o_t^i, define the active clipping mask* $m_t^i = \begin{cases} 1, & A_t^i \geq 0 \text{ and } r_t^i \leq 1+\epsilon, \\ 1, & A_t^i < 0 \text{ and } r_t^i \geq 1-\epsilon, \\ 0, & \text{otherwise} \end{cases}$ *and the detached token weight $w_t^i = \text{sg}(m_t^i r_t^i)$, where $\text{sg}(\cdot)$ denotes the stop-gradient operator. Consider the weighted log-likelihood objective in Eq. (4), then we have that $\nabla_\theta\mathcal{J}_{\text{GRPO}}(\theta) = \nabla_\theta\mathcal{J}_{\text{WLL}}(\theta)$.*

$$\mathcal{J}_{WLL}(\theta) = \mathbb{E}_{q\sim\mathcal{D},\{o^i\}_{i=1}^G\sim\pi_{\theta_{old}}(\cdot|q)} \left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o^i|}\sum_{t=1}^{|o^i|}\left(w_t^i A_t^i \log\pi_\theta(o_t^i|q,o_{<t}^i) - \beta\mathbb{D}_{KL}(\pi_\theta\|\pi_{ref})\right)\right]. \tag{4}$$

Proof. We prove in three cases:

Case 1: $A_t^i \geq 0$ and $r_t^i > 1+\epsilon$. We have

$$\mathcal{J}_{\text{WLL}}(\theta) = \mathbb{E}_{q\sim\mathcal{D},\{o^i\}_{i=1}^G\sim\pi_{\theta_{\text{old}}}(\cdot|q)} \left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o^i|}\sum_{t=1}^{|o^i|}\left(-\beta\mathbb{D}_{\text{KL}}(\pi_\theta\|\pi_{\text{ref}})\right)\right]. \tag{5}$$

Also, we have

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim\mathcal{D},\{o^i\}_{i=1}^G\sim\pi_{\theta_{\text{old}}}(\cdot|q)} \left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o^i|}\sum_{t=1}^{|o^i|}\left((1+\varepsilon)\hat{A}_t^i - \beta\mathbb{D}_{\text{KL}}(\pi_\theta\|\pi_{\text{ref}})\right)\right]. \tag{6}$$

Thus, we have $\nabla_\theta\mathcal{J}_{\text{GRPO}}(\theta) = \nabla_\theta\mathcal{J}_{\text{WLL}}(\theta)$.

Case 2: $A_t^i < 0$ and $r_t^i < 1-\epsilon$. We have

$$\mathcal{J}_{\text{WLL}}(\theta) = \mathbb{E}_{q\sim\mathcal{D},\{o^i\}_{i=1}^G\sim\pi_{\theta_{\text{old}}}(\cdot|q)} \left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o^i|}\sum_{t=1}^{|o^i|}\left(-\beta\mathbb{D}_{\text{KL}}(\pi_\theta\|\pi_{\text{ref}})\right)\right]. \tag{7}$$

Also, we have

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim\mathcal{D},\{o^i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q)} \left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o^i|}\sum_{t=1}^{|o^i|}\left((1-\varepsilon)\hat{A}_t^i - \beta\mathbb{D}_{\text{KL}}(\pi_\theta\|\pi_{\text{ref}})\right)\right]. \quad (8)$$

Thus, we have $\nabla_\theta\mathcal{J}_{\text{GRPO}}(\theta) = \nabla_\theta\mathcal{J}_{\text{WLL}}(\theta)$.

Case 3: Otherwise, we have

$$\mathcal{J}_{\text{WLL}}(\theta) = \mathbb{E}_{q\sim\mathcal{D},\{o^i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q)} \left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o^i|}\sum_{t=1}^{|o^i|}\left(sg\left(\frac{\pi_\theta(o_t^i|q,o_{<t}^i)}{\pi_{old}(o_t^i|q,o_{<t}^i)}\right)A_t^i\log\pi_\theta(o_t^i|q,o_{<t}^i) - \beta\mathbb{D}_{\text{KL}}(\pi_\theta\|\pi_{\text{ref}})\right)\right]. \quad (9)$$

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim\mathcal{D},\{o^i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q)} \left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o^i|}\sum_{t=1}^{|o^i|}\left(\frac{\pi_\theta(o_t^i|q,o_{<t}^i)}{\pi_{old}(o_t^i|q,o_{<t}^i)}\hat{A}_t^i - \beta\mathbb{D}_{\text{KL}}(\pi_\theta\|\pi_{\text{ref}})\right)\right]. \quad (10)$$

Thus, we have $\nabla_\theta\mathcal{J}_{\text{GRPO}}(\theta) = \nabla_\theta\mathcal{J}_{\text{WLL}}(\theta)$.

□

B.2 Gradient Analysis

The gradient of original GRPO is as follows:

$$\begin{aligned}
\nabla_\theta\mathcal{J}(\theta) &= \nabla_\theta\mathbb{E}_{q\sim\mathcal{D},\{o^i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q)}\left[\frac{1}{G}\sum_{i=1}^{G}r_t^i(\theta)\hat{A}_t^i\right] \\
&= \mathbb{E}_{q\sim\mathcal{D},\{o^i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q)}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{\nabla_\theta\pi_\theta(o_t^i\mid q,o_{<t}^i)}{\pi_{\theta_{\text{old}}}(o_t^i\mid q,o_{<t}^i)}\hat{A}_t^i\right] \\
&= \mathbb{E}_{q\sim\mathcal{D},\{o^i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q)}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{\pi_\theta(o_t^i\mid q,o_{<t}^i)}{\pi_{\theta_{\text{old}}}(o_t^i\mid q,o_{<t}^i)}\nabla_\theta\log\pi_\theta(o_t^i\mid q,o_{<t}^i)\hat{A}_t^i\right]
\end{aligned} \quad (11)$$

where π_θ denotes $\pi_\theta(o_t^i\mid q,o_{<t}^i)$ and $\pi_{\theta_{\text{old}}}$ denotes $\pi_{\theta_{\text{old}}}(o_t^i\mid q,o_{<t}^i)$. Then, we derive the gradient of positive tokens in ASPO as follows:

$$\begin{aligned}
\nabla_\theta\mathcal{J}(\theta) &= \nabla_\theta\mathbb{E}_{q\sim\mathcal{D},\{o^i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q)}\left[\frac{1}{G}\sum_{i=1}^{G}\hat{r}_t^i(\theta)\hat{A}_t^i\right] \\
&= \mathbb{E}_{q\sim\mathcal{D},\{o^i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q)}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{\pi_{\theta_{\text{old}}}(o_t^i\mid q,o_{<t}^i)\nabla_\theta\pi_\theta(o_t^i\mid q,o_{<t}^i)}{\text{sg}(\pi_\theta^2(o_t^i\mid q,o_{<t}^i))}\hat{A}_t^i\right] \\
&= \mathbb{E}_{q\sim\mathcal{D},\{o^i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q)}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{\pi_{\theta_{\text{old}}}(o_t^i\mid q,o_{<t}^i)\pi_\theta(o_t^i\mid q,o_{<t}^i)}{\pi_\theta^2(o_t^i\mid q,o_{<t}^i)}\nabla_\theta\log\pi_\theta(o_t^i\mid q,o_{<t}^i)\hat{A}_t^i\right] \\
&= \mathbb{E}_{q\sim\mathcal{D},\{o^i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q)}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{\pi_{\theta_{\text{old}}}(o_t^i\mid q,o_{<t}^i)}{\pi_\theta(o_t^i\mid q,o_{<t}^i)}\nabla_\theta\log\pi_\theta(o_t^i\mid q,o_{<t}^i)\hat{A}_t^i\right]
\end{aligned} \quad (12)$$

It is worth noting that the gradient of positive tokens in equation 12 differs from the original gradient of GRPO in equation 11 at the highlighted red term. From the above derivation, we can observe that the gradient of ASPO is positively correlated with $\frac{1}{\pi_\theta}$, indicating that the gradient becomes larger when the probability of a token is lower.

C Experimental Details

Baselines. We conduct experiments using DeepSeek-R1-Distill-Qwen-1.5B (Guo et al., 2025), Qwen3-4B, Qwen3-8B, and Qwen3-30B-A3B (Yang et al., 2025a) as the base model and compare ASPO with the following baselines:

- **Base Model**: The original model without any RL fine-tuning.
- **GRPO** (Shao et al., 2024): A strong OSRL algorithm with token-level loss.
- **DeepScaleR-1.5B** (Luo et al., 2025b): A 1.5B model trained for mathematical reasoning with iterative context-length expansion.
- **DeepCoder-1.5B** (Luo et al., 2025a): A 1.5B model trained on code datasets using similar context expansion strategies as DeepScaleR.
- **Nemotron-1.5B** (Liu et al., 2025a): A strong 1.5B reasoning model with reference policy resetting.

Evaluation. We follow standard practice (Liu et al., 2025a; Wang et al., 2025) and report both *avg@K* and *pass@K* metrics. For benchmarks with fewer samples (AIME24/25 and AMC23), we set $K = 64$. For LiveCodeBench v6, we use $K = 16$; for LiveCodeBench v5 and Minerva Math, $K = 8$; and for MATH-500 and OlympiadBench, $K = 4$. To ensure fair and accurate evaluation, we adopt the official verification functions from both DeepScaleR and Math-Verify[2] for mathematical problems, following the protocol in He et al. (2025).

Implementation Details. For training data, we use a mixture of DeepScaleR-Preview-Dataset (Luo et al., 2025b), Skywork-OR1-RL-Data (He et al., 2025), and DAPO-Math-17K (Yu et al., 2025) for DeepSeek-R1-Distill-Qwen-1.5B on mathematical tasks. For Qwen3-4B, Qwen3-8B, and Qwen3-30B-A3B, we directly use DAPO-Math-17K for training. For coding, we employ DeepCoder (Luo et al., 2025a), CodeContests (Li et al., 2022), and CodeForces (Penedo et al., 2025) datasets. All datasets are cleaned and filtered following the preprocessing protocol of Wang et al. (2025). After filtering, the mathematical dataset contains 70.8k samples, while the coding dataset contains 8.9k samples. The clipping ranges of GRPO and ASPO are set to $\varepsilon = 0.2$. The KL divergence is incorporated as an explicit loss term (k3 estimator) with coefficient $\beta = 0.001$, and all baselines use the same KL implementation to ensure that observed differences in entropy and stability are not caused by KL. For each prompt, 16 responses are sampled with a temperature of 1.0. We use a max response length of 32,768 tokens for DeepSeek-R1-Distill-Qwen-1.5B, 8,192 for Qwen3-4B and Qwen3-30B-A3B, and 6,144 for Qwen3-8B. Experiments of 1.5B, 4B, and 8B LLMs are conducted on 8 NVIDIA H800 GPUs and Qwen3-30B-A3B is trained with 32 NVIDIA H800 GPUs.

D Additional Experimental Results

We present the training dynamics and test performance curves on mathematical tasks in Figure 5, 6, and 7. As shown in Figures 5 and 6, although ASPO exhibits slightly slower learning during the initial training phase compared to GRPO, it maintains more stable model entropy and consistently lower repetition rate throughout training. These results demonstrate the superior training stability of ASPO.

Ablation on Dual Clipping. Weight Flipping is the core mechanism of ASPO, while Dual Clipping is a stabilizer introduced because, after inversion, extreme positive-token ratios shift to the side where their magnitude also needs to be bounded. As shown in Figure 8, without Dual Clipping, the gradient norm exhibits explosive spikes during training, while ASPO with Dual Clipping maintains stable gradient dynamics. This confirms that Dual Clipping is essential for stabilizing training in extreme cases, complementing the Weight Flipping mechanism.

E Interpreting Training Stability and Premature Collapse

To better contextualize our analysis, we introduce an important concept in RL training: how to distinguish between **healthy convergence** and **local optima**.

In RL training, healthy convergence typically exhibits the following characteristics:

[2] https://github.com/huggingface/Math-Verify

Table 3: Evaluation results on OOD benchmarks. The results of ASPO are shaded and the highest values are **bolded**.

Method	ARC-Challenge	GPQA-Diamond	MMLU-Pro	Avg.
Qwen3-4B	74.7	14.3	31.5	40.2
↳ GRPO	**90.1**	29.2	49.6	56.3
↳ ASPO-Math-4B	87.6	**37.3**	**53.5**	**59.5**
Qwen3-8B	52.9	8.6	29.6	30.4
↳ GRPO	62.4	33.0	42.9	46.1
↳ ASPO-Math-8B	**88.6**	**56.3**	**57.4**	**67.4**

Table 4: Evaluation results with Qwen3-4B. The highest values are **bolded**.

Method	AIME24	AIME25	AMC23	MATH-500	Minerva	Olympiad	Avg.
GRPO	43.4	35.5	84.3	91.7	47.2	**67.4**	61.6
GRPO w/o IS	46.4	36.3	**88.9**	**92.9**	**49.3**	67.2	63.5
GRPO w/ Pos. Resp. IS Mean	**46.5**	**42.3**	87.4	92.0	48.6	67.1	**64.0**

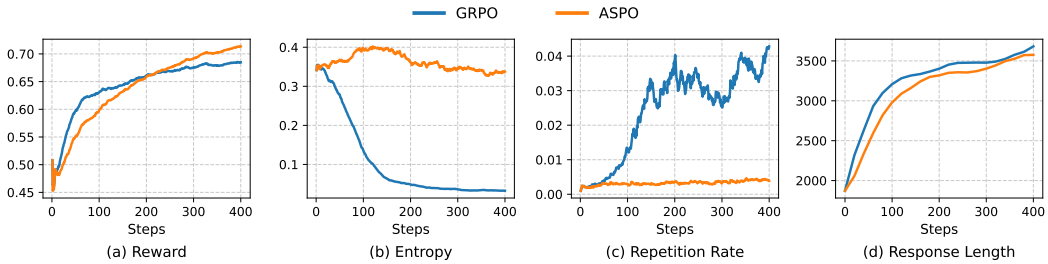


Figure 5: Training dynamics of GRPO and ASPO on Qwen3-4B, including (a) training reward, (b) model entropy, (c) repetition rate, and (d) response length. The curves are smoothed with EMA for better visualization.

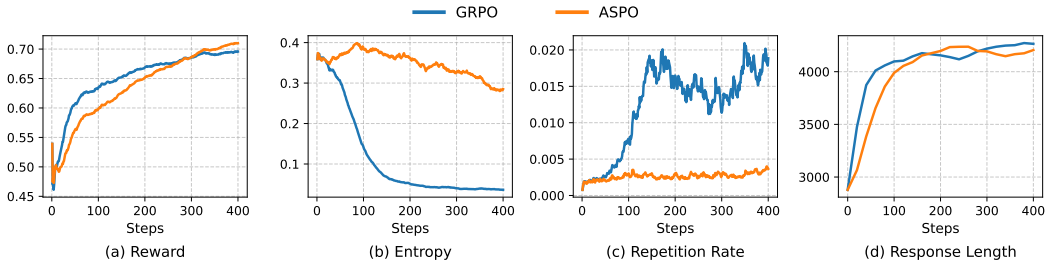


Figure 6: Training dynamics of GRPO and ASPO on Qwen3-8B, including (a) training reward, (b) model entropy, (c) repetition rate, and (d) response length. The curves are smoothed with EMA for better visualization.

- **Entropy** decreases gradually from a relatively high initial value and stabilizes at a small but positive level, indicating that the policy becomes more deterministic while retaining moderate exploration.
- The **reward curve** increases steadily and eventually plateaus at a stable high value.
- **Clip ratios** and **KL divergence loss** remain stable during later training stages, without drastic fluctuations.

In contrast, when the training becomes trapped in a local optimum, the model enters a self-reinforcing **policy-data distribution loop**, characterized by:

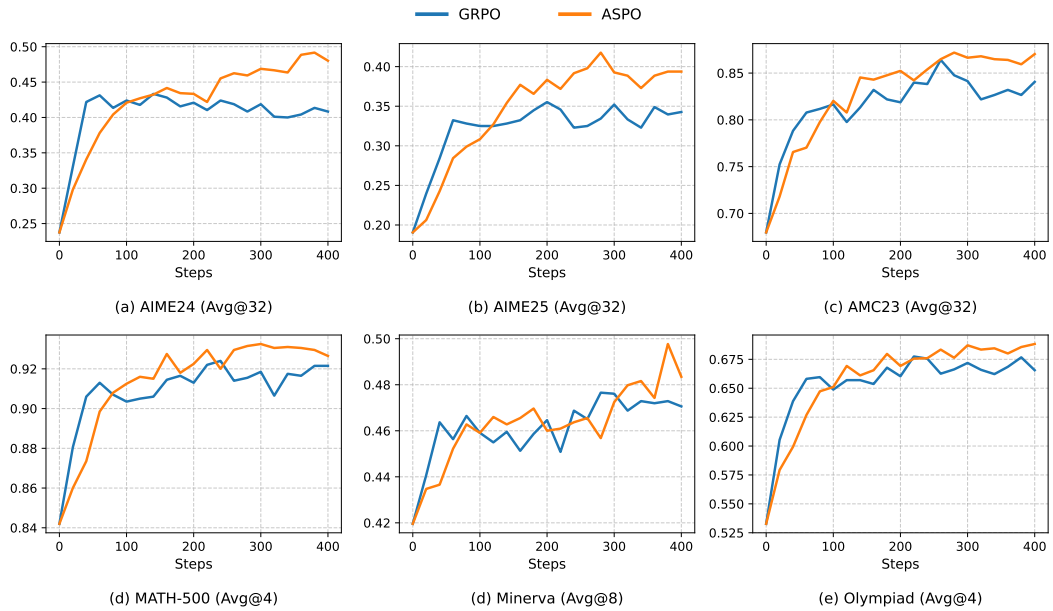


Figure 7: The test curves of all methods trained with Qwen3-4B on six mathematical benchmarks.

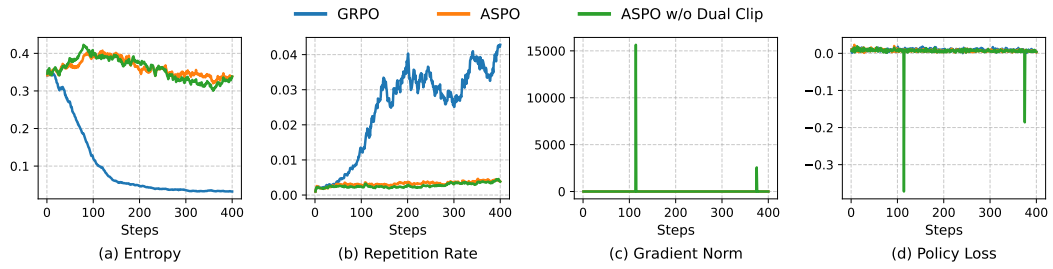


Figure 8: Ablation on dual clipping. The training dynamics curves of ASPO with and without dual clipping on Qwen3-4B, including (a) entropy, (b) repetition rate, (c) gradient norm, and (d) policy loss. Without dual clipping, the gradient norm and policy loss exhibit explosive spikes during training, demonstrating that dual clipping is essential for stabilizing training in extreme cases. The curves are smoothed with EMA for better visualization.

- **Entropy** collapses rapidly toward zero.
- The **reward curve** stagnates with no further improvement.
- Persistently high **clip ratios** without meaningful policy updates.

When training GRPO-based approaches, we may observe this phenomenon: in late-stage training, entropy collapses, repetition rates spike, clip ratios surge, and the model's test performance begins to degrade, suggesting premature convergence.

In early training, **moderate entropy reduction** and a **gradual increase in clip ratio** are expected. If accompanied by an **increasing rewards**, it indicates that the policy is learning effectively. However, when later stages exhibit an abrupt entropy drop, reward stagnation, and persistently high clip ratios without further progress, it suggests **premature convergence**.

The underlying cause lies in the **token-level weight mismatch for positive samples in PPO-Clip** identified in Section 4, which drives the model to **overfit** certain high-probability tokens, eventually leading to entropy collapse and training degradation.