SPURIOUS POLICY SHIFT: DO NOT BE TOO CONSERVATIVE FOR YOUR POLICY OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning (RL) has become a key component of post-training for large language models (LLMs), where token-level clipping stabilizes off-policy optimization by controlling token-wise importance sampling (IS) ratios between the current policy and the rollout policy. However, we empirically identify a phenomenon termed *spurious policy shift* (SPS): a large shift in the sampled token's probability may simply reflect probability mass shifting among contextually interchangeable alternatives at the same decoding state, rather than a meaningful change in the model's generation policy. Furthermore, we observe that SPS-affected tokens occupy a substantial portion of upper-clipped positive advantage updates and suppress policy gain, suggesting that the standard token-level clipping is too conservative and may unnecessarily suppress positive policy updates. Building on this observation, we propose SPS-aware clipping for policy optimization (SCPO) that identifies SPS through probability mass shifting between the sampled token and its contextually interchangeable candidate tokens under the old and new policies, and selectively relaxes clipping for such tokens to avoid suppressing positive policy updates. Experimental results show that our method outperforms state-of-the-art methods on multiple math reasoning benchmarks without compromising training stability. Our code is available at: https://anonymous.4open.science/r/scpo.

1 INTRODUCTION

Reinforcement learning (RL) has become the dominant paradigm for post-training large language models (LLMs), driving substantial gains on complex tasks such as mathematical reasoning, coding, and agentic decision-making (Ouyang et al., 2022; Jaech et al., 2024; Guo et al., 2025; Comanici et al., 2025). In current RL-based LLM post-training pipelines, PPO (Schulman et al., 2017) and GRPO (Shao et al., 2024) style methods typically stabilize policy optimization by clipping token-level importance ratios when the new policy changes the probability of sampled tokens too aggressively (Schulman et al., 2017; Shao et al., 2024; Yu et al., 2025). Recent clipping variants further show that the choice of clipping range can affect entropy, optimization balance, and off-policy stability (Yang et al., 2025b; Xi et al., 2026). However, this clipping mechanism relies on an implicit assumption: a large token-level probability shift is taken as an evidence of a large and potentially risky policy shift.

In this work, we empirically show that this assumption can be overly conservative for policy optimization, suppressing a subset of postive model updates. Unlike actions in conventional RL, token choices in LLMs are often locally structured: at many decoding positions, the sampled token has several contextually interchangeable alternatives. As shown in Figure 1, a large shift in the sampled token's probability may simply reflect the redistribution of probabilities among locally similar candidate tokens, rather than a meaningful change in the model's generation policy. We term this phenomenon *spurious policy shift* (SPS). From this perspective, standard token-level clipping may misinterpret harmless local probability redistribution as an overly large policy update.

More importantly, our empirical analysis shows that SPS-affected tokens occupy a substantial portion of upper-clipped positive advantage updates and account for many of the positive updates clipped away by standard clipping during RL training. This observation suggests that the standard token-level clipping may be over-conservative in RL training: it may suppress positive policy

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.

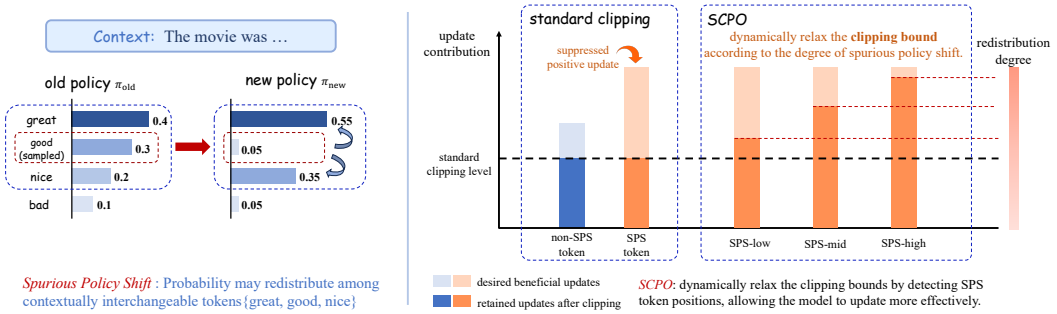


Figure 1: Policy optimization exhibits spurious policy shifts (SPS). SCPO addresses SPS by dynamically relaxing clipping bounds for SPS tokens.

updates that correspond only to the redistribution of probabilities within a contextually interchangeable tokens. This issue may reduce the effective learning step and limit the degree of policy update, even when the update remains locally well-behaved.

Building on the above observation, we propose **SPS-aware Clipping for Policy Optimization (SCPO)**. SCPO identifies SPS-affected tokens by comparing the probability change of the sampled token with the change in the total probability assigned to its contextually interchangeable alternatives under the old and new policies. As shown in Figure 1, SCPO selectively relaxes the *upper clipping bound* according to the degree of SPS exhibited by sampled tokens, so that positive policy updates are less likely to be suppressed, while standard clipping is retained for non-spurious risky policy shifts.

Experimental results show that SCPO consistently outperforms state-of-the-art methods on multiple math reasoning benchmarks without compromising training stability. These findings support a simple but important conclusion: in policy optimization, token-level probability shift is not always a reliable proxy for meaningful policy shift, and clipping should account for local probability redistribution among SPS-affected tokens. Our work makes three contributions:

1. We identify *spurious policy shift* (SPS) in RL training of LLM policy optimization, and we further show that SPS-affected tokens account for a substantial fraction of upper-clipped positive advantage updates and suppress policy gain.
2. We propose **SCPO**, an SPS-aware clipping method for LLM policy optimization that detects SPS by comparing probability redistribution under the old and new policies, and dynamically relaxes the *upper clipping bound* for SPS-affected tokens to reduce unnecessary suppression of positive policy updates while preserving training stability.
3. We validate SCPO across multiple backbones and evaluation benchmarks, showing that it consistently improves RL performance while maintaining stable optimization.

2 Preliminaries

Before introducing spurious policy shift (SPS) and our method, we first review the basic policy gradient formulation for large language model (LLM) reinforcement learning (RL) and the clipped policy optimization objective commonly used in GRPO-style training (Schulman et al., 2017; Shao et al., 2024). These preliminaries establish the notation for token-level states, advantages, and importance-sampling ratios.

Policy Gradient for LLM RL. Given an input prompt $\mathbf{x}$ sampled from a training set $\mathcal{D}$, a language model policy π_θ autoregressively generates a response $\mathbf{y} = (y_1, \ldots, y_T)$ according to

$$\pi_\theta(\mathbf{y} \mid \mathbf{x}) = \prod_{t=1}^{T} \pi_\theta(y_t \mid \mathbf{x}, \mathbf{y}_{<t}), \tag{1}$$

where $\mathbf{y}_{<t} = (y_1, \ldots, y_{t-1})$ denotes the prefix before step t. For convenience, we denote the decoding state at step t by $s_t = (\mathbf{x}, \mathbf{y}_{<t})$. Let $R(\mathbf{x}, \mathbf{y})$ denote the reward assigned to the generated

response. The standard RL objective for LLM post-training is to maximize the expected reward:

$$J(\theta) = \mathbb{E}_{\mathbf{x}\sim\mathcal{D},\,\mathbf{y}\sim\pi_\theta(\cdot|\mathbf{x})}\left[R(\mathbf{x},\mathbf{y})\right]. \tag{2}$$

Policy gradient methods optimize this objective by weighting the log-probability of each generated token with an advantage signal:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\mathbf{x}\sim\mathcal{D},\,\mathbf{y}\sim\pi_\theta(\cdot|\mathbf{x})}\left[\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta(y_t \mid s_t)\, A_t\right], \tag{3}$$

where A_t denotes the advantage at step t. In practice, A_t may be instantiated by different estimators used in LLM RL, such as GRPO style token-level advantages. Our method is orthogonal to the specific choice of advantage estimator and focuses instead on how the policy update is clipped once A_t is given.

Clipped Policy Optimization. In LLM RL training, data are typically generated by a rollout policy π_{old}, while optimization is performed on the current policy π_θ. To account for the mismatch between the two policies, clipped policy optimization introduces the token-level importance sampling ratio

$$r_t(\theta) = \frac{\pi_\theta(y_t \mid s_t)}{\pi_{\text{old}}(y_t \mid s_t)} = \exp(\log \pi_\theta(y_t \mid s_t) - \log \pi_{\text{old}}(y_t \mid s_t)). \tag{4}$$

Given token-level advantages A_t, a widely used GRPO-style clipped surrogate objective can be written as

$$\mathcal{L}_{\text{clip}}(\theta) = \mathbb{E}\left[\sum_{t=1}^{T}\min(r_t(\theta)A_t,\ \text{clip}(r_t(\theta),\,1-\epsilon_{\text{low}},\,1+\epsilon_{\text{high}})\,A_t)\right], \tag{5}$$

where ϵ_{low} and ϵ_{high} control the clipping range. This objective constrains the magnitude of model's policy updates by suppressing tokens whose probability ratios move beyond the prescribed interval.

A key property of this formulation is that the clipping decision is made solely according to the probability ratio of the *sampled token* y_t. In other words, the magnitude of token-level probability shift is directly used as the proxy for whether the policy update at that position is overly large. This design is effective in practice and underlies many GRPO style LLM RL methods. At the same time, it also raises a central question for our work: does a large shift in the sampled token's probability always indicate a meaningful and risky change in the model's policy behavior?

3 Spurious Policy Shift

In this section, we identify and analyze *spurious policy shift* (SPS), a phenomenon in policy optimization where an apparent large shift in sampled-token probability may reflect probability redistribution among contextual candidates, rather than a genuine change in the model's policy. Specifically, we first define a probability-based contextual candidate set and then formalize SPS. Furthermore, we empirically show that ***SPS tokens account for a substantial fraction of upper-clipped positive advantage updates and the policy gain removed by clipping, revealing a conservative behavior of standard token-level clipping.***

3.1 Definition of Spurious Policy Shift

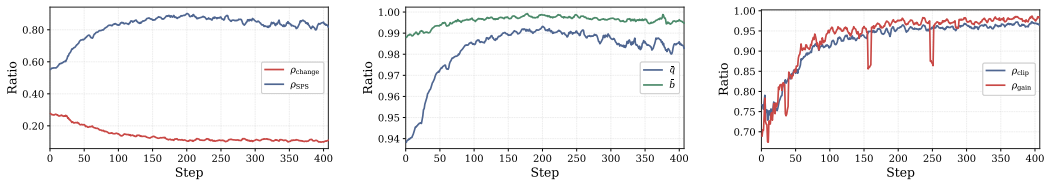
We first formalize spurious policy shift as a structured relation between the probability change of a sampled token and the probability change of its contextual candidates. Consider a response $\mathbf{y}$ and let $\mathcal{R}$ denote the set of valid response positions. For each $i \in \mathcal{R}$, the decoding state is $\mathbf{s}_i = (\mathbf{x}, \mathbf{y}_{<i})$ and the sampled token is $y_i \in \mathcal{V}$, where $\mathcal{V}$ denotes the vocabulary. Let π_{old} denote the rollout policy and π_θ denote the current policy.

Contextually interchangeable candidates. For each token $v \in \mathcal{V} \setminus \{y_i\}$, we use its rollout-policy probability under the same decoding state as the candidate score:

$$\kappa_i(v) = \pi_{\text{old}}(v \mid \mathbf{s}_i). \tag{6}$$

Let $(v_{i,1}, \ldots, v_{i,|\mathcal{V}|-1})$ be an ordering of $\mathcal{V} \setminus \{y_i\}$ such that

$$\kappa_i(v_{i,1}) \geq \kappa_i(v_{i,2}) \geq \cdots \geq \kappa_i(v_{i,|\mathcal{V}|-1}), \tag{7}$$



(a) (b) (c)

Figure 2: Analysis of spurious policy shift during RL training. (a) The overall rate of large changed probability tokens decreases, while the SPS rate among changed tokens remains high. (b) Changed tokens exhibit strong counter-balancing and local conservation. (c) SPS occupies a large fraction of upper-clipped positive-gain updates and accounts for a large share of the suppressed gain.

with ties broken by a fixed deterministic rule. The contextual candidate set is then defined as

$$\mathcal{N}_i = \{v_{i,1}, \ldots, v_{i,K}\}. \tag{8}$$

Thus, $\mathcal{N}_i$ contains the K most likely alternatives to y_i under the rollout policy at the same decoding state. Such candidates often serve as contextually interchangeable continuations, so probability redistribution within $\mathcal{N}_i$ need not imply a genuine change in the model's policy. We fix $\mathcal{N}_i$ under π_{old} before comparing it with π_θ, ensuring that candidate-mass changes are measured over the same contextual candidate set.

Evaluation metrics of SPS. We use two scalar quantities to compare π_{old} and π_θ at position i. The first is the probability assigned to the sampled token y_i: $p_i^{\text{old}} = \pi_{\text{old}}(y_i \mid \mathbf{s}_i)$ and $p_i^\theta = \pi_\theta(y_i \mid \mathbf{s}_i)$. The sampled-token shift is then defined as $\Delta_i^{\text{tok}} = p_i^\theta - p_i^{\text{old}}$. The second quantity is the total probability mass assigned to the contextual candidate set $\mathcal{N}_i$: $m_i^{\text{old}} = \sum_{v \in \mathcal{N}_i} \pi_{\text{old}}(v \mid \mathbf{s}_i)$ and $m_i^\theta = \sum_{v \in \mathcal{N}_i} \pi_\theta(v \mid \mathbf{s}_i)$. The contextual-candidate mass shift is then defined as $\Delta_i^{\text{cand}} = m_i^\theta - m_i^{\text{old}}$. To determine whether a sampled-token shift is induced by redistribution within the contextual candidate set, we introduce two structural scores. The first score measures *counter-balancing*. If probability mass is redistributed between the sampled token and its contextual candidates, then an increase in the sampled-token probability should be accompanied by a decrease in candidate mass, and vice versa. We therefore define the counter-balancing score as

$$b_i = \frac{\left[-\operatorname{sign}\left(\Delta_i^{\text{tok}}\right) \Delta_i^{\text{cand}}\right]_+}{\left|\Delta_i^{\text{tok}}\right| + \varepsilon}, \tag{9}$$

where $[z]_+ = \max(z, 0)$ and $\varepsilon > 0$ is a small constant. The numerator keeps only the component of candidate-mass change that offsets the sampled-token shift, and the denominator normalizes it by the magnitude of the sampled-token shift. Thus, a larger b_i means that the apparent token-level change is more strongly counter-balanced by contextual candidates. However, counter-balancing alone is not sufficient: a candidate set may move in the opposite direction while the combined probability mass still changes substantially. We therefore introduce a second score, *local conservation*, which measures whether the total probability mass over the sampled token and its contextual candidates remains approximately stable:

$$q_i = 1 - \frac{\left|\Delta_i^{\text{tok}} + \Delta_i^{\text{cand}}\right|}{\left|\Delta_i^{\text{tok}}\right| + \left|\Delta_i^{\text{cand}}\right| + \varepsilon}. \tag{10}$$

Definition of SPS. For a response position $i \in \mathcal{R}$, we say that the sampled token y_i exhibits *spurious policy shift* if it satisfies three conditions: a sufficiently large apparent sampled-token shift, sufficient counter-balancing from contextual candidates, and sufficient local conservation. Formally, the SPS set is

$$\mathcal{S} = \left\{ i \in \mathcal{R} \;\middle|\; \left|\Delta_i^{\text{tok}}\right| \geq \tau,\ b_i \geq \gamma,\ q_i \geq \eta \right\}. \tag{11}$$

Here, τ controls the minimum apparent sampled-token shift, while γ and η control the required strengths of counter-balancing and local conservation, respectively.

3.2 Existence and Impact of Spurious Policy Shift

Existence of SPS. We next quantify whether SPS is a persistent phenomenon during RL training. The analysis focuses on positions where the sampled-token probability changes substantially:

$$\mathcal{C} = \left\{ i \in \mathcal{R} \;\middle|\; \left|\Delta_i^{\text{tok}}\right| \geq \tau \right\}. \tag{12}$$

This set represents the apparent large token-level shifts observed by the clipping rule. We first examine whether these large shifts exhibit the structural pattern required by SPS. To this end, we report the average counter-balancing $\bar{b}$ and local-conservation $\bar{q}$ scores over $\mathcal{C}$:

$$\bar{b} = \frac{1}{|\mathcal{C}|} \sum_{i \in \mathcal{C}} b_i, \qquad \bar{q} = \frac{1}{|\mathcal{C}|} \sum_{i \in \mathcal{C}} q_i. \tag{13}$$

These quantities measure whether large sampled-token shifts are, on average, accompanied by candidate-side compensation and local mass conservation. We then measure the scale of the phenomenon:

$$\rho_{\text{change}} = \frac{|\mathcal{C}|}{|\mathcal{R}|}, \qquad \rho_{\text{SPS}} = \frac{|\mathcal{S}|}{|\mathcal{C}|}. \tag{14}$$

Here, ρ_{change} measures how frequently substantial sampled-token shifts occur, while ρ_{SPS} measures the fraction of such shifts that satisfy the SPS definition.

Figure 2(a) shows a clear training pattern. The analysis is conducted on GRPO. The contextual candidate set is constructed with $K = 4$, and SPS is identified with $\tau = 0.005$, $\gamma = 0.995$, and $\eta = 0.995$. As RL training progresses, ρ_{change} steadily decreases, indicating that large sampled-token probability change become less frequent. However, among the remaining changed tokens, ρ_{SPS} increases and remains high. Figure 2(b) further shows that $\bar{b}$ and $\bar{q}$ stay close to their upper range under the same analysis criterion, indicating strong candidate-side compensation and local mass conservation. These observations suggest that in the later stage of training, although the overall update magnitude becomes smaller, the remaining large token-level shifts are dominated by contextual redistribution rather than genuine changes in policy behavior.

Impact of SPS. Although we empirically show that SPS exists during RL training, this alone does not show that SPS changes the optimization process. We therefore examine the part of the clipped objective where useful learning signal can be removed. For each response position $i \in \mathcal{R}$, define the token-level importance sampling ratio r_i and let A_i denote its advantage. When $A_i > 0$, increasing r_i increases the unclipped objective contribution $A_i r_i$. However, if r_i exceeds the upper clipping boundary, standard clipping caps this contribution at $A_i(1+\epsilon_{\text{high}})$. We denote these capped positive-gain tokens by

$$\mathcal{U}^+ = \{ i \in \mathcal{R} \mid A_i > 0,\ r_i > 1 + \epsilon_{\text{high}} \}, \tag{15}$$

where ϵ_{high} is the upper clipping width. Although the SPS definition itself is symmetric with respect to increases and decreases in sampled-token probability, the following analysis focuses on positive advantage positions of SPS. This is the case directly affected by upper clipping, and relaxing both upper and lower clipping boundaries may harm training stability, which we will return to this design choice in the experimental ablations. Here, we measure how many upper-clipped positive advantage positions are SPS positions:

$$\rho_{\text{clip}} = \frac{|\mathcal{S} \cap \mathcal{U}^+|}{|\mathcal{U}^+|}. \tag{16}$$

A large ρ_{clip} means that a substantial fraction of the updates truncated by upper clipping are locally redistributive SPS updates.

This count-based statistic does not capture how much positive objective value is removed. For each $i \in \mathcal{U}^+$, the unclipped contribution is $A_i r_i$, while the clipped contribution is $A_i(1 + \epsilon_{\text{high}})$. The amount removed by upper clipping is therefore

$$g_i^{\text{sup}} = A_i \left(r_i - (1 + \epsilon_{\text{high}}) \right), \qquad i \in \mathcal{U}^+. \tag{17}$$

We then measure the fraction of suppressed gain coming from SPS positions:

$$\rho_{\text{gain}} = \frac{\sum_{i \in \mathcal{S} \cap \mathcal{U}^+} g_i^{\text{sup}}}{\sum_{i \in \mathcal{U}^+} g_i^{\text{sup}}}. \tag{18}$$

Figure 2(c) shows that SPS is heavily concentrated in the updates truncated by upper clipping, using the same SPS identification setting as (a) and (b). Both ρ_{clip} and ρ_{gain} remain high throughout training, and ρ_{gain} is generally comparable to or higher than ρ_{clip}. This indicates that SPS not only occupies a large fraction of capped positive-gain updates, but also accounts for a large share of the positive objective value removed by clipping.

Together, these results reveal a late-stage mismatch in token-level clipping. As training progresses, large token-level shifts become rarer, but the remaining large shifts are increasingly dominated by SPS. Standard clipping still treats these redistributive shifts as risky whenever their sampled-token ratios exceed the fixed boundary, thereby removing a substantial portion of the already limited positive update signal. This motivates a clipping mechanism that relaxes unnecessary truncation for SPS while preserving clipping for genuine policy shifts.

4 SPS-AWARE CLIPPING FOR POLICY OPTIMIZATION (SCPO)

In this section, we propose **SPS-aware Clipping for Policy Optimization (SCPO)**, a new policy optimization method that accounts for spurious policy shifts in LLM reinforcement learning. SCPO estimates the strength of SPS at each token and uses it to adaptively relax the upper clipping boundary, so that locally redistributive beneficial updates are less likely to be prematurely truncated.

4.1 SPS-AWARE RELAXATION INTENSITY

Standard clipping uses a fixed importance sampling (IS) ratio range for all tokens, and therefore cannot distinguish SPS from genuine policy shifts. Relaxing the range by a fixed amount for all SPS tokens is still insufficient: a binary SPS label only indicates that the token satisfies the minimum SPS criteria, but does not measure how strongly it exhibits counter-balancing and local conservation. A fixed relaxation would therefore either over-relax weak SPS cases or under-relax strong SPS cases. SCPO instead uses a continuous SPS intensity to determine a token-specific clipping range. We first define the binary SPS indicator

$$M_i = \mathbb{I}[i \in \mathcal{S}], \tag{19}$$

where $\mathcal{S}$ is the SPS set defined in Eq. equation 11. To measure the strength of SPS beyond the decision boundary, we use the margins by which the two structural scores exceed their thresholds. For the counter-balancing score b_i, we define

$$s_i^{(b)} = 1 - \exp\left(-\frac{[b_i - \gamma]_+}{\max(\gamma, \varepsilon)}\right), \tag{20}$$

where $[z]_+ = \max(z, 0)$. This score is zero when $b_i \le \gamma$ and increases smoothly as the counter-balancing evidence becomes stronger. For the local-conservation score q_i, we define

$$s_i^{(q)} = \mathrm{clip}\left(\frac{[q_i - \eta]_+}{\max(1 - \eta, \varepsilon)}, 0, 1\right), \tag{21}$$

where $\mathrm{clip}(z, a, b) = \min(\max(z, a), b)$. This score is zero when $q_i \le \eta$ and approaches one as local conservation becomes stronger.

The final SPS relaxation intensity is

$$\alpha_i = M_i \cdot \frac{s_i^{(b)} + s_i^{(q)}}{2}, \qquad \alpha_i \in [0, 1]. \tag{22}$$

Thus, non-SPS positions have $\alpha_i = 0$, while SPS positions receive larger α_i when they show stronger counter-balancing and local conservation. This intensity is the quantity used by SCPO to determine the token-specific clipping range. During optimization, α_i serves only as a per-step adaptive clipping coefficient. We stop gradients through the SPS detector and α_i, so the policy gradient flows only through $r_i(\theta)$, with SPS statistics recomputed after each policy update.

4.2 SPS-AWARE POLICY OPTIMIZATION

We now define the SCPO optimization rule. The central design is to replace the fixed upper clipping boundary with a token-dependent boundary determined by the SPS relaxation intensity. For a response position i, let

$$r_i(\theta) = \frac{\pi_\theta(y_i \mid \mathbf{s}_i)}{\pi_{\mathrm{old}}(y_i \mid \mathbf{s}_i)} \tag{23}$$

be the token-level IS ratio, and let A_i denote the corresponding advantage. In our experiments, A_i is instantiated by the GRPO advantage.

SCPO uses α_i to define a token-dependent upper clipping width:

$$\epsilon_{\text{high},i}^{\text{SCPO}} = \epsilon_{\text{high}} \left(1 + \mathbb{I}[A_i > 0]\, \alpha_i^{\rho}(\lambda - 1)\right), \tag{24}$$

where ϵ_{high} is the base upper clipping width, $\lambda \geq 1$ is the maximum relaxation multiplier, and $\rho > 0$ controls how sharply the relaxation grows with α_i. The indicator $\mathbb{I}[A_i > 0]$ ensures that the relaxation applies only to beneficial probability-increasing updates. The lower clipping width remains fixed at ϵ_{low}. Thus, the clipped ratio under SCPO is

$$\bar{r}_i^{\text{SCPO}}(\theta) = \text{clip}\left(r_i(\theta), 1 - \epsilon_{\text{low}}, 1 + \epsilon_{\text{high},i}^{\text{SCPO}}\right). \tag{25}$$

Compared with standard clipping, the only change is that the upper boundary becomes token-dependent and SPS-aware. Thus, SCPO maximizes the following policy objective:

$$\mathcal{J}_{\text{SCPO}}(\theta) = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\pi_{\text{old}}}\left[\frac{1}{|\mathcal{R}|}\sum_{i\in\mathcal{R}} \min\left(r_i(\theta)A_i,\, \bar{r}_i^{\text{SCPO}}(\theta)A_i\right)\right]. \tag{26}$$

This objective reduces to the standard clipped objective whenever $\alpha_i = 0$. When $\alpha_i > 0$ and $A_i > 0$, SCPO relaxes the upper clipping boundary in proportion to the estimated SPS intensity, allowing locally redistributive positive updates to contribute more fully. At the same time, clipping is still applied, so updates without SPS evidence remain constrained by the original clipping range.

5 Experiments

In this section, we evaluate SCPO on challenging mathematical reasoning benchmarks. We use Qwen3-4B (Yang et al., 2025a) as the backbone model and keep the training and evaluation settings consistent across all compared methods. Evaluation results for larger-scale models are detailed in the Appendix C.1.

5.1 Setup

Model and training data. We use Qwen3-4B as the base model for all experiments. Unless otherwise specified, all methods share the same model initialization, tokenizer, rollout pipeline, reward function, and evaluation protocol. For RL fine-tuning, we follow the DAPO training setup (Yu et al., 2025), using the same mathematical reasoning corpus and verifier-based reward setting.

Evaluation benchmarks and metrics. We evaluate mathematical reasoning performance on six benchmarks, including AIME'24, AIME'25, AMC'23, MATH500, Minerva, and Olympiad-Bench (Hendrycks et al., 2021; Lewkowycz et al., 2022; He et al., 2024). For each test problem, we sample 16 candidate solutions and report two metrics: **avg@16**, the average accuracy over all sampled solutions, and **pass@16**, the fraction of problems for which at least one of the 16 samples is correct. Higher values indicate better reasoning performance.

Baselines. We compare SCPO with representative policy optimization methods under the same backbone and evaluation protocol. The baselines include token-level GRPO (Shao et al., 2024) and DAPO (Yu et al., 2025), sequence-level GSPO (Zheng et al., 2025) and dynamic clipping range methods such as DCPO (Yang et al., 2025b). These comparisons evaluate whether SPS-aware clipping improves policy optimization beyond existing token-level and sequence-level stabilization strategies.

Implementation details. For a fair comparison, all methods use the same training configuration for all methods: a training batch size of 32, a minibatch size of 4, and a group size of 8. All methods use the same sampling strategy during evaluation and generate 16 candidate solutions per problem for computing avg@16 and pass@16. For SCPO, the contextual candidate set is constructed from the top-K ($K = 4$) rollout-policy candidates at each response position. Unless otherwise stated, the lower clipping width is kept unchanged, and SCPO only adapts the upper clipping boundary according to the SPS relaxation intensity.

5.2 Main Results

SCPO achieves overall superior performance. Table 1 presents the main results on mathematical reasoning benchmarks. Under the same Qwen3-4B training setting, SCPO obtains the best average

Table 1: Main results on mathematical reasoning benchmarks using Qwen3-4B. We report avg@16 and pass@16. Best results are in **bold**, and second-best results are underlined.

Method	AIME'24		AIME'25		AMC'23		MATH500		Minerva		OlympiadBench		Average	
	avg@16	pass@16	avg@16	pass@16	avg@16	pass@16	avg@16	pass@16	avg@16	pass@16	avg@16	pass@16	avg@16	pass@16
GRPO	18.54	40.00	15.00	26.67	60.47	80.00	76.94	90.40	35.64	50.74	39.45	56.46	41.01	57.38
DAPO	22.71	50.00	15.83	36.67	59.69	85.00	78.56	91.60	28.91	45.96	42.25	59.44	41.32	61.44
GSPO	27.12	46.67	**26.04**	**43.33**	**74.06**	92.50	77.45	89.80	36.08	49.63	40.90	56.32	46.94	63.04
DCPO	26.88	50.00	22.92	40.00	71.88	92.50	84.69	93.20	37.32	46.69	47.50	60.18	48.53	63.76
SCPO	**27.92**	**56.67**	25.00	40.00	73.75	**95.00**	**84.91**	**94.60**	**37.87**	**51.84**	**48.77**	**64.49**	**49.70**	**67.10**

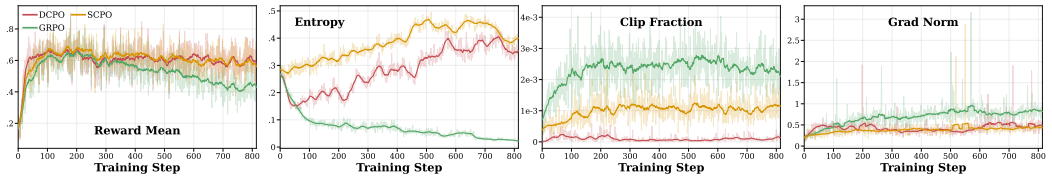


Figure 3: Training dynamics of different policy optimization methods. From left to right, we report reward mean, entropy, clip fraction, and gradient norm over training steps.

performance across benchmarks and remains within the top two on every reported metric. Specifically, SCPO achieves the best avg@16 on four of six benchmarks and the best pass@16 on five of six benchmarks. GSPO performs strongly on AIME'25 and AMC'23, where it obtains the best avg@16 scores, but SCPO maintains stronger overall performance across the full benchmark suite. On average, SCPO improves over GRPO by 8.70 points in avg@16 and 9.72 points in pass@16, outperforms DCPO by 1.17 points in avg@16 and 3.34 points in pass@16, and exceeds GSPO by 2.76 points in avg@16 and 4.06 points in pass@16. These results show that SCPO consistently improves policy optimization by preserving beneficial updates that are unnecessarily suppressed by standard token-level clipping.

Training Dynamics. Final benchmark performance alone does not show how a method behaves during RL training. We therefore compare the training dynamics of SCPO with GRPO and DCPO. Figure 3 reports four training curves: reward mean, entropy, clip fraction, and gradient norm. The specific meanings of these metrics are shown in Appendix D.2

Reward mean. SCPO achieves a stronger and more stable reward trajectory than GRPO. While all methods improve rapidly at the beginning, GRPO starts to decline after the early stage, indicating that standard clipping becomes less effective at sustaining policy improvement. SCPO maintains a higher reward level throughout the later training stage, suggesting that SPS-aware relaxation preserves useful positive updates without destabilizing reward optimization.

Entropy. SCPO maintains the highest entropy among the compared methods. GRPO exhibits a rapid entropy collapse, while DCPO recovers entropy gradually but remains below SCPO for most of training. This indicates that SCPO does not obtain its gains by making the policy prematurely over-confident. Instead, by relaxing clipping only for SPS-supported updates, SCPO improves optimization efficiency while preserving a broader output distribution.

Clip fraction. SCPO keeps a substantially lower clip fraction than GRPO while remaining higher than the nearly flat DCPO curve. This behavior is consistent with the design of SCPO: it reduces unnecessary clipping for locally redistributive updates, but does not remove clipping globally. In contrast, GRPO maintains a much higher clip fraction, suggesting that many updates continue to be truncated under the fixed token-level clipping rule.

Gradient norm. SCPO shows a stable gradient norm throughout training. Its curve stays close to DCPO and below GRPO for most training steps, without sustained spikes or growth. This suggests that dynamically relaxing the upper clipping boundary for SPS tokens does not introduce unstable gradients. Together with the reward and entropy curves, this shows that SCPO improves update efficiency while maintaining stable policy optimization.

5.3 Ablation Study

In this section, we ablate the three design choices that are specific to SCPO: using contextual probability candidates to detect SPS, relaxing only the upper clipping bound for positive advantage SPS updates, and removing the upper clipping constraint.

Table 2: Ablation study of SCPO on mathematical reasoning benchmarks. We report avg@16 and pass@16. Higher is better. Best results are in **bold**.

Method	AIME'24		AIME'25		AMC'23		MATH500		Minerva		OlympiadBench	
	avg@16	pass@16	avg@16	pass@16	avg@16	pass@16	avg@16	pass@16	avg@16	pass@16	avg@16	pass@16
SCPO	**27.92**	**56.67**	**25.00**	**40.00**	73.75	**95.00**	**84.91**	94.60	**37.87**	**51.84**	**48.77**	**64.49**
w/ embedding	15.21	33.33	17.71	33.33	63.75	87.50	82.38	93.40	35.82	47.43	44.11	57.21
w/ lower relaxation	16.67	46.67	12.50	36.67	50.47	90.00	66.90	91.40	32.31	48.53	31.51	56.61
w/o upper clipping	27.50	43.33	23.96	**40.00**	**77.03**	92.50	84.61	**94.80**	37.34	**51.84**	47.82	63.60

Embedding-neighbor candidate set. SCPO assumes that SPS is a contextual redistribution effect, so the neighboring candidates should be tokens that compete under the same decoding state. Here, we examine whether embedding similarity can serve as a proxy for semantic similarity by replacing the rollout-policy top-K candidate set with a static embedding-neighbor set:

$$\mathcal{N}_i^{\text{emb}} = \text{TopK}_{v \in \mathcal{V} \setminus \{y_i\}} \cos(\mathbf{e}_v, \mathbf{e}_{y_i}), \tag{27}$$

where $\mathbf{e}_v$ denotes the token embedding of token v. All other components of SCPO are kept unchanged.

Upper-and-lower relaxation. SCPO relaxes only the upper clipping bound because the SPS effect we target appears in upper-clipped positive advantage updates. Since the SPS definition itself is symmetric with respect to probability increases and decreases, we further test a symmetric variant that keeps the upper clipping width the same as Eq. equation 24 and additionally relaxes the lower clipping width:

$$\epsilon_{\text{low},i}^{\text{sym}} = \epsilon_{\text{low}} \left(1 + \mathbb{I}[A_i < 0] \alpha_i^{\rho} (\lambda - 1)\right). \tag{28}$$

The corresponding objective is

$$\mathcal{J}_{\text{sym}}(\theta) = \mathbb{E}_{(\mathbf{x},\mathbf{y}) \sim \pi_{\text{old}}} \left[\frac{1}{|\mathcal{R}|} \sum_{i \in \mathcal{R}} \min \left(r_i(\theta) A_i,\ \text{clip} \left(r_i(\theta), 1 - \epsilon_{\text{low},i}^{\text{sym}}, 1 + \epsilon_{\text{high},i}^{\text{SCPO}} \right) A_i \right) \right]. \tag{29}$$

Without upper clipping. To test whether SCPO's gains simply come from allowing larger positive advantage updates, we consider a variant that removes the upper clipping constraint. This variant uses

$$\bar{r}_i^{\text{no-upper}}(\theta) = \max \left(r_i(\theta), 1 - \epsilon_{\text{low}} \right), \tag{30}$$

and optimizes the same clipped objective with $\bar{r}_i^{\text{no-upper}}(\theta)$ in place of $\bar{r}_i^{\text{SCPO}}(\theta)$. Thus, probability increases are no longer capped, while the lower clipping boundary is kept unchanged. This ablation tests whether selective SPS-aware relaxation is necessary, or whether simply removing the upper bound is sufficient.

Results. Table 2 shows that all three design choices are important. Replacing contextual probability candidates with embedding neighbors reduces the average performance by 6.54 points in avg@16 and 8.40 points in pass@16, confirming that SPS is better captured by context-dependent candidate competition than by static embedding similarity. Relaxing both upper and lower clipping bounds leads to an even larger drop of 14.64 points in avg@16 and 5.45 points in pass@16, supporting our focus on upper-clipped positive-advantage updates. Removing upper clipping altogether achieves a similar average avg@16 to SCPO, but lowers pass@16 by 2.75 points, suggesting that the gains do not come from globally eliminating the upper bound. Instead, SCPO benefits from selectively relaxing SPS updates while preserving clipping for updates without SPS evidence.

6 Conclusion

In this work, we identify a phenomenon termed *spurious policy shift* (SPS) in large language models' policy optimization, where large sampled-token probability shifts may reflect probability redistribution among contextual candidates rather than genuine policy changes. We show that SPS is prevalent during reinforcement learning (RL) training and accounts for a substantial portion of upper-clipped positive-advantage updates and the corresponding suppressed policy gain. Building on this observation, we propose SCPO, an SPS-aware policy optimization method that estimates SPS intensity and dynamically relaxes the upper clipping boundary for SPS tokens. Experiments on mathematical reasoning benchmarks demonstrate that SCPO improves RL performance while maintaining stable training dynamics, and ablation studies validate the importance of contextual candidate construction and upper-only relaxation.

AI USE STATEMENT

In this research, generative AI tools were employed as part of the methodology, specifically for experiments involving the training, evaluation, and analysis of large language models (LLMs). These tools were also used to assist with refining language, correcting grammar, and improving the clarity and readability of the manuscript. We did not use generative AI tools to independently generate or determine the scientific conclusions of this work, and the other disclosure categories do not apply to this study. All AI-assisted text was reviewed and revised by the authors to ensure technical accuracy, consistency with the experimental results, and proper attribution. The authors independently verified the experimental results, analyses, and claims presented. We accept full responsibility for the final content of this work, including all text, claims, analyses, and other materials created using generative AI.

REFERENCES

Charles Arnal, Gaëtan Narozniak, Vivien Cabannes, Yunhao Tang, Julia Kempe, and Remi Munos. Asymmetric REINFORCE for off-policy reinforcement learning: Balancing positive and negative rewards. In *Proceedings of The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Xin Zhao, Zhenliang Zhang, and Furu Wei. Reasoning with exploration: An entropy perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

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

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, et al. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*, 2025.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 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, 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*, 2024.

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.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. In *Proceedings of the 35th Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Aitor Lewkowycz, Anders Johan Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Venkatesh 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 *Proceedings of the Thirty-Sixth Conference on Neural Information Processing Systems*, 2022.

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. In *Second Conference on Language Modeling*, 2025.

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. In *Proceedings of the Thirty-Sixth Conference on Neural Information Processing Systems*, 2022.

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

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

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

Qwen Team. Qwq-32b: Embracing the power of reinforcement learning, 2025.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xiong-Hui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for LLM reasoning. In *Proceedings of The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Zhiheng Xi, Xin Guo, Yang Nan, Enyu Zhou, Junrui Shen, Wenxiang Chen, Jiaqi Liu, Jixuan Huang, Xun Deng, Zhihao Zhang, Honglin Guo, Zhikai Lei, Miao Zheng, Guoteng Wang, Peng Sun, Rui Zheng, Hang Yan, Tao Gui, Qi Zhang, and Xuanjing Huang. BAPO: Stabilizing off-policy reinforcement learning for LLMs via balanced policy optimization with adaptive clipping. In *The Fourteenth International Conference on Learning Representations*, 2026.

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*, 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.

Zhihe Yang, Xufang Luo, Zilong Wang, Dongqi Han, Zhiyuan He, Dongsheng Li, and Yunjian Xu. Do not let low-probability tokens over-dominate in rl for llms. *arXiv preprint arXiv:2505.12929*, 2025c.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, YuYue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, LingJun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Yonghui Wu, and Mingxuan Wang. DAPO: An open-source LLM reinforcement learning system at scale. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Kaiyan Zhang, Yuxin Zuo, Bingxiang He, Youbang Sun, Runze Liu, Che Jiang, Yuchen Fan, Kai Tian, Guoli Jia, Pengfei Li, et al. 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.

APPENDIX

Algorithm 1 SCPO: SPS-aware Clipping for Policy Optimization

Require: Policy π_θ, dataset $\mathcal{D}$, reward function R, group size G, candidate size K, SPS thresholds (τ, γ, η), clipping widths $(\epsilon_{\text{low}}, \epsilon_{\text{high}})$, relaxation hyperparameters (λ, ρ)
Ensure: Updated policy π_θ

for training step $s = 1, \ldots, S$ **do**
  Sample responses:
  Set rollout policy $\pi_{\text{old}} \leftarrow \pi_\theta$
  Sample prompts $\mathbf{x} \sim \mathcal{D}$ and responses $\{\mathbf{y}^{(g)}\}_{g=1}^{G} \sim \pi_{\text{old}}(\cdot \mid \mathbf{x})$
  Compute rewards and token-level advantages $\{A_i\}$ using R
  Estimate SPS intensity:
  for each valid response position $i \in \mathcal{R}$ **do**
    Build contextual candidate set $\mathcal{N}_i$ from top-K rollout-policy candidates
    Compute sampled-token shift Δ_i^{tok} and candidate-mass shift Δ_i^{cand}
    Compute counter-balancing score b_i and local-conservation score q_i
    $M_i \leftarrow \mathbb{I}\big[|\Delta_i^{\text{tok}}| \geq \tau,\ b_i \geq \gamma,\ q_i \geq \eta\big]$
    Compute relaxation intensity α_i from M_i, b_i, and q_i according to Eq. equation 22
  end for
  Dynamic SPS-aware clipping:
  for each valid response position $i \in \mathcal{R}$ **do**
    Compute token-level ratio $r_i(\theta)$
    Set $\epsilon_{\text{high},i}^{\text{SCPO}}$ according to Eq. equation 24
    Set $\bar{r}_i^{\text{SCPO}}(\theta) \leftarrow \text{clip}\Big(r_i(\theta),\ 1-\epsilon_{\text{low}},\ 1+\epsilon_{\text{high},i}^{\text{SCPO}}\Big)$
  end for
  Update policy:
  Update π_θ by maximizing $\mathcal{J}_{\text{SCPO}}(\theta)$ in Eq. equation 26
end for

A ALGORITHM

A.1 THE PSEUDO-CODE OF SCPO

Algorithm 1 summarizes the training procedure of SCPO at the token level. For each sampled response, SCPO first constructs a contextual candidate set for every valid token using the top-K rollout-policy probabilities. It then compares the rollout policy and the current policy on both the sampled token and its contextual candidates, obtaining the sampled-token shift and the contextual-candidate mass shift. Based on these two shifts, SCPO computes the counter-balancing score and the local-conservation score, which determine whether the token exhibits SPS and how strong the SPS structure is. The resulting SPS intensity is used to construct a token-dependent upper clipping width, while the lower clipping width remains unchanged. Finally, the policy is updated with the SCPO objective in Eq. equation 26. This design preserves the standard clipped policy optimization form while selectively relaxing upper clipping for locally redistributive positive updates.

B RELATED WORK

Reinforcement learning for reasoning language models. Recent reasoning-oriented models, including OpenAI o1 (Jaech et al., 2024), DeepSeek-R1 (Guo et al., 2025), Gemini 2.5 (Comanici et al., 2025), and QwQ (Team, 2025), show that reinforcement learning can substantially improve long-chain reasoning, mathematical problem solving, and agentic decision-making in large language models. This line of work builds on policy-gradient training for language models, with PPO-style objectives and GRPO-style group-relative advantages becoming common choices for verifier-guided reasoning optimization (Schulman et al., 2017; Shao et al., 2024; Zhang et al., 2025). In these methods, token-level likelihood ratios provide a practical bridge between sampled trajectories and policy updates, while clipping prevents overly aggressive changes. Our work follows this clipped policy

optimization paradigm, but asks whether the sampled-token ratio alone is always a faithful signal of meaningful policy shift in LLM generation.

Clipping, entropy, and training stability. Although clipped policy optimization has enabled strong LLM RL results, recent studies show that stability remains a central bottleneck. DAPO (Yu et al., 2025) improves large-scale RL through techniques such as Clip-Higher and dynamic sampling, while GSPO (Zheng et al., 2025) reformulates optimization at the sequence level. Other work analyzes how entropy evolves during reasoning RL, showing that entropy collapse, low-probability token dominance, and insufficient exploration can limit further improvement (Cui et al., 2025; Wang et al., 2025; Cheng et al., 2026; Liu et al., 2025; He et al., 2025; Yang et al., 2025c). These studies motivate algorithms that treat clipping not only as a generic trust-region heuristic, but also as a mechanism that shapes which token updates can contribute to learning.

Adaptive and off-policy clipping. Several recent methods modify the clipping rule itself. Asymmetric REINFORCE and tapered off-policy REINFORCE study off-policy learning with separate treatment of positive and negative rewards or tapered importance weights (Arnal et al., 2025; Roux et al., 2025). DCPO adapts token-level clipping according to token prior probabilities, aiming to reduce harmful domination by low-probability tokens (Yang et al., 2025b). BAPO analyzes the imbalance between positive and negative optimization contributions in off-policy RL and introduces adaptive clipping bounds to improve entropy preservation and training stability (Xi et al., 2026). These methods adjust clipping according to reward-sign asymmetry, importance-weight tapering, token prior probability, or aggregate optimization balance.

C Additional Experiments

C.1 Larger Scaling on Qwen3-8B

Table 3 reports the scaling results using Qwen3-8B as the backbone. SCPO achieves the best avg@16 on all six benchmarks and the best or tied-best pass@16 on five out of six benchmarks, showing that SPS-aware clipping remains effective at a larger model scale. On average, SCPO improves over GRPO by 2.56 points in avg@16 and 4.35 points in pass@16, and outperforms DCPO by 4.47 points in avg@16 and 2.85 points in pass@16. Notably, DCPO underperforms GRPO on several benchmarks, including AIME'24, AIME'25, MATH500, and OlympiadBench in avg@16, suggesting that dynamically adjusting both lower and upper clipping bounds can introduce unstable optimization behavior. This is consistent with our ablation in Table 2, where additionally relaxing the lower bound substantially degrades performance. In contrast, SCPO focuses on SPS-supported upper-bound relaxation, preserving the benefit of adaptive clipping while avoiding the instability caused by broad two-sided relaxation.

Table 3: Scaling results on mathematical reasoning benchmarks using Qwen3-8B. We report avg@16 and pass@16. Higher is better. Best results are in **bold**.

Method	AIME'24		AIME'25		AMC'23		MATH500		Minerva		OlympiadBench	
	avg@16	pass@16	avg@16	pass@16	avg@16	pass@16	avg@16	pass@16	avg@16	pass@16	avg@16	pass@16
GRPO	33.96	56.75	27.08	40.73	81.41	93.77	88.40	93.73	41.11	50.65	51.03	61.33
DCPO	31.67	60.00	25.63	40.00	80.47	**95.00**	84.51	93.80	41.27	**55.15**	48.50	62.26
SCPO	**37.08**	**66.72**	**30.00**	**51.73**	**85.16**	**95.00**	**90.05**	**94.41**	**41.54**	50.44	**54.50**	**65.39**

D Experimental settings

D.1 The details of experiments

For the Math task, uniform hyperparameters are employed across all methods. The maximum response lengths are set to 4096 tokens, respectively. The actor learning rate is fixed at 1.0e-6. We employ group-based rollouts with a group size of 8. A binary rule-based reward function is used (1 for success, 0 for failure). The batch sizes for training and evaluation are 32 and 64, respectively. During evaluation, we use nucleus sampling with top_p = 1.0 and a temperature of 1.0. For comparing methods, we use the suggested parameters in their papers. The parameters of SCPO are the

same as those presented in Section 3.2. Experiments of Qwen3-4B are conducted on NVIDIA H100 GPUs and Qwen3-8B are conducted on NVIDIA H200 GPUs.

D.2 THE MEANINGS OF TRAINING METRICS

- *Gradient Norm*: Measures the magnitude of policy gradients during optimization. Stable and moderate values suggest controlled updates; sharp spikes or persistently large norms indicate overly aggressive updates and potential training instability.
- *Policy Gradient Clip Fraction*: Fraction of updates affected by clipping. Moderate levels suggest controlled policy steps; persistently high fractions imply frequent extreme ratio deviations and unstable optimization dynamics.
- *Entropy*: Tracks policy stochasticity. Adequate entropy supports exploration and prevents premature collapse; rapid entropy decay indicates mode collapse or over-confident generation, while excessively high entropy can slow convergence.
- *Mean Reward*: Average return per rollout. A smooth upward trend reflects effective learning; abrupt drops typically signal instability or reward hacking.