Soft Sequence Policy Optimization

Svetlana Glazyrina[1], **Maksim Kryzhanovskiy**[1,2], **Roman Ischenko**[1,2]

[1]Lomonosov Moscow State University, Moscow, Russia
[2]Institute for Artificial Intelligence, Lomonosov Moscow State University, Moscow, Russia
glazyrinase@my.msu.ru, kryzhanovskiyms@my.msu.ru, rois@cs.msu.ru

Abstract

A significant portion of recent research on Large Language Model (LLM) alignment focuses on developing new policy optimization methods based on Group Relative Policy Optimization (GRPO). Two prominent directions have emerged: (i) a shift toward sequence-level importance sampling weights that better align with the sequence-level rewards used in many tasks, and (ii) alternatives to the PPO-style clipping that aim to avoid the associated loss of training signal and entropy collapse. We introduce Soft Sequence Policy Optimization, an off-policy reinforcement learning objective that incorporates soft gating functions over token-level probability ratios within sequence-level importance weights. We provide theoretical motivation for SSPO and investigate practical modifications to improve optimization behavior. Empirically, we demonstrate that SSPO improves training stability and performance both in mathematical reasoning and coding tasks.

1 Introduction

Reinforcement learning (RL) has become a central ingredient in enhancing the reasoning abilities of large language models (LLMs), particularly for tasks requiring long chains of thought (CoT), multi-step decision making, and delayed rewards. While supervised fine-tuning excels at imitating local patterns present in static datasets, RL enables optimization over entire generated sequences, making it especially attractive for complex reasoning, program synthesis, and decision-centric language modeling.

Among modern RL approaches for LLM alignment, group-based policy optimization methods have emerged as a practical and effective strategy. These methods sample multiple candidate responses per input prompt and normalize sequence-level rewards within each group, yielding a reliable prompt value estimate without relying on an auxiliary critic network. Representative algorithms such as Group Relative Policy Optimization (GRPO) (Shao et al., 2024) and REINFORCE Leave-One-Out (RLOO) (Ahmadian et al., 2024) demonstrate that relative comparisons within a group can significantly stabilize training, reduce computational overhead, and improve reasoning performance across a wide range of domains.

Despite their empirical success, existing group-based methods face important limitations when deployed at scale. In particular, off-policy learning becomes unavoidable in realistic training pipelines. As model sizes increase, architectures incorporate sparsity (e.g., Mixture-of-Experts), and generated sequences become longer, large rollout batches are required to fully utilize modern hardware. To improve sample efficiency, these batches are typically partitioned into multiple mini-batches for gradient updates. This practice inevitably induces an off-policy setting, where responses are sampled from a behavior policy while updates are applied to a newer policy, a regime in which GRPO-like algorithms increasingly dominate current practice.

In off-policy group-based optimization, policy updates are usually weighted by importance sampling (IS) ratios between the current and behavior policies. These ratios are well known to suffer from high variance, especially for long sequences where token-level likelihood ratios compound multiplicatively. Uncontrolled IS weights can destabilize training and ultimately degrade performance. Two broad strategies have emerged to mitigate this issue.

First, many methods apply hard clipping to large importance weights, limiting their influence on the gradient (Schulman et al., 2017; Shao et al., 2024; MiniMax et al., 2025). While effective at reducing variance, hard clipping introduces a difficult tradeoff: aggressive clipping improves stability but reduces sample efficiency and limits exploration (Chen et al., 2023; Dwyer et al., 2025), whereas loose clipping preserves learning signal at the cost

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 53191–53201
October 24-29, 2026 ©2026 Association for Computational Linguistics

of noisy and brittle updates.

Second, in accordance with insights from RLOO and motivated by the prevalence of sequence-level rewards in LLM alignment (e.g., RLHF and RLVR), recent work proposes enforcing the sequence-level coherence of importance weights rather than treating token-level contributions independently. This perspective motivates objectives such as Group Sequence Policy Optimization (GSPO) (Zheng et al., 2025) and Geometric-Mean Policy Optimization (GMPO) (Zhao et al., 2025), which replace arithmetic aggregation of token-level ratios with multiplicative or geometric formulations that better respect the structure of sequence probabilities. A key innovation of GMPO is the combination of sequence-level weighting with token-level probability ratio clipping.

However, existing solutions remain incomplete. Sequence-coherent objectives, such as GSPO, improve stability but do not fully address the interaction between off-policy learning and entropy-regularized objectives commonly used in modern RL for LLMs. In parallel, soft policy optimization methods such as Soft Adaptive Policy Optimization (SAPO) (Gao et al., 2025) highlight the importance of entropy-aware objectives that smoothly interpolate between exploitation and exploration. However, they are not explicitly designed for sequence-level optimization.

In this paper, we propose *Soft Sequence Policy Optimization (SSPO)*, a new off-policy reinforcement learning objective that unifies and extends insights from reward coherent sequence weights and soft policy optimization. SSPO introduces a soft sequence-level weighting mechanism with token-level weight attenuation that controls importance sampling variance without resorting to hard clipping while preserving coherent credit assignment to entire responses. By operating at the sequence level and incorporating entropy-aware regularization, SSPO achieves a more favorable exploration-exploitation tradeoff than prior approaches in off-policy group-based RL.

Our contributions are as follows:

- We analyze the properties of token-wise gate functions when importance weights are aggregated via the geometric mean.
- We propose Soft Sequence Policy Optimization (SSPO), a sequence-level objective with soft clipping for off-policy reinforcement learning.
- We demonstrate improved training stability and performance on various model sizes, architectures, and domains.

2 Preliminaries

Since we do not modify the additive KL-divergence regularization term in the loss, we omit it for brevity.

Notation Following standard conventions in the literature, an autoregressive language model parameterized by θ is defined by a policy π_θ. x denotes a query and $\mathcal{D}$ denotes the set of queries. Given a response y to a query x, the likelihood of y under the policy π_θ is defined as

$$\pi_\theta(y \mid x) = \prod_{t=1}^{|y|} \pi_\theta(y_t \mid x, y_{<t}),$$

where $|y|$ denotes the number of tokens in y. Each query-response pair (x, y) is assigned a reward $r(x, y)$ by a verifier r.

Throughout this paper, we adopt a compact notation for advantage-dependent clipping operators commonly used in policy optimization objectives. Rather than explicitly writing expressions of the form

$$\min\Big(\rho \cdot \widehat{A},\ \mathrm{clip}(\rho,\ 1-\varepsilon_{\text{low}},\ 1+\varepsilon_{\text{high}}) \cdot \widehat{A}\Big),$$

we define the advantage-aware clipping function

$$f_{\text{Clip}}(\rho; \widehat{A}) \triangleq \begin{cases} \min(\rho, 1+\varepsilon_{\text{high}}), & \widehat{A} > 0, \\ \max(\rho, 1-\varepsilon_{\text{low}}), & \widehat{A} \le 0, \end{cases} \quad (1)$$

and write the clipped weights compactly as $f_{\text{Clip}}(\rho;\ \widehat{A})\ \cdot \widehat{A}$.

This notation enables a unified presentation of hard and soft clipping mechanisms in the subsequent sections.

Group Relative Policy Optimization Building on Proximal Policy Optimization (PPO) (Schulman et al., 2017), Group Relative Policy Optimization (GRPO) (Shao et al., 2024) enables off-policy learning by leveraging samples generated by an older (behavior) policy $\pi_{\theta_{\text{old}}}$. Following PPO, GRPO employs a clipping mechanism to constrain policy updates within a proximal region around the behavior policy. The key difference is that GRPO replaces computationally and memory-intensive advantage estimators based on a critic network, whose reliability may be questionable (Kazemnejad et al.,

2025), with a relative advantage computed among responses to the same query. The GRPO optimization objective is

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\,\{y_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|x)} \left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y_i|}\sum_{t=1}^{|y_i|} f_{\text{Clip}}(\rho_{i,t}(\theta);\,\widehat{A}_{i,t})\cdot\widehat{A}_{i,t}\right]. \quad (2)$$

Here, G denotes the group size, i.e., the number of responses generated for a single query. The importance ratio $\rho_{i,t}(\theta)$ and the advantage $\widehat{A}_{i,t}$ for token $y_{i,t}$ are defined as

$$\rho_{i,t}(\theta) = \frac{\pi_\theta(y_{i,t}|x, y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t}|x, y_{i,<t})}, \quad (3)$$

$$\widehat{A}_{i,t} = \widehat{A}_i = \frac{r(x, y_i) - \text{mean}\left(\{r(x, y_i)\}_{i=1}^{G}\right)}{\text{std}\left(\{r(x, y_i)\}_{i=1}^{G}\right)}. \quad (4)$$

The notation $\widehat{A}_i = \widehat{A}_{i,t}$ is valid since all tokens in the sequence y_i share the same advantage.

Group Sequence Policy Optimization Group Sequence Policy Optimization (GSPO) (Zheng et al., 2025) identifies a primary problem of GRPO: a mismatch between the unit of optimization and the unit of reward. In GRPO, importance sampling weights for off-policy correction and clipping are applied at the token level, whereas the advantage remains constant across all tokens in a sequence. In contrast, GSPO performs optimization, and in particular clipping, directly at the sequence level. Its objective is given by

$$\mathcal{J}_{\text{GSPO}}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\,\{y_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|x)} \left[\frac{1}{G}\sum_{i=1}^{G} f_{\text{Clip}}(s_i(\theta);\,\widehat{A}_i)\cdot\widehat{A}_i\right], \quad (5)$$

where the importance ratio $s_i(\theta)$ is computed based on the sequence likelihood, with length normalization to control variance (Zheng et al., 2023):

$$s_i(\theta) = \left(\frac{\pi_\theta(y_i|x)}{\pi_{\theta_{\text{old}}}(y_i|x)}\right)^{\frac{1}{|y_i|}} \quad (6)$$

GSPO has been shown to improve training stability for large models and to increase sample efficiency compared to GRPO, although the fraction of clipped tokens grows.

Geometric-Mean Policy Optimization Geometric-Mean Policy Optimization (GMPO) (Zhao et al., 2025) combines response-level importance sampling with token-level probability ratio clipping. GMPO adopts the same sequence-length normalization as GSPO but is motivated by a different intuition: reducing variance in token-level importance sampling weights and sensitivity to outliers by replacing the arithmetic mean over tokens in the GRPO objective with a geometric mean. Specifically, GMPO optimizes the following objective:

$$\mathcal{J}_{\text{GMPO}}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\{y_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|x)} \left[\frac{1}{G}\sum_{i=1}^{G}\left[\prod_{t=1}^{|y_i|}\left|f_{\text{Clip}}(\rho_{i,t}(\theta);\widehat{A}_i)\right|\right]^{\frac{1}{|y_i|}}\cdot\hat{A}_i\right]. \quad (7)$$

This geometric aggregation improves robustness to tokens with disproportionately large importance weights and reduces the range of the loss values, leading to more stable updates. The authors further observe that token-level clipping is less aggressive than sequence-level clipping and more effectively constrains the range of importance sampling ratios.

Soft Adaptive Policy Optimization In existing group-based policy optimization methods, high variance in importance sampling weights is typically mitigated through hard clipping. This introduces a trade-off between update stability and sample efficiency and may also reduce exploration abilities. As an alternative, Soft Adaptive Policy Optimization (SAPO) (Gao et al., 2025) employs a smooth, temperature-controlled gating mechanism that preserves the learning signal for all sampled tokens. The SAPO objective is defined as:

$$\mathcal{J}_{\text{SAPO}}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\{y_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|x)} \left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y_i|}\sum_{t=1}^{|y_i|} f_{\text{Soft}}(\rho_{i,t}(\theta);\,\widehat{A}_{i,t})\cdot\widehat{A}_{i,t}\right], \quad (8)$$

where

$$f_{\text{Soft}}(x;\widehat{A}) = \sigma\left(\tau(\widehat{A})\cdot(x-1)\right)\cdot\frac{4}{\tau(\widehat{A})}, \quad (9)$$

$$\tau(\widehat{A}) = \begin{cases}\tau_{\text{pos}}, & \text{if } \widehat{A} > 0\\ \tau_{\text{neg}}, & \text{otherwise}\end{cases}. \quad (10)$$

Here, τ_{pos} and τ_{neg} denote the temperatures for positive and negative advantages, respectively, and $\sigma(x) = 1/(1+e^{-x})$ is the sigmoid function.

By construction, the sigmoid-shaped gating function preserves gradients during on-policy updates while remaining bounded. Attenuating the contribution of outlier tokens, rather than truncating it entirely, allows SAPO to maintain informative learning signals throughout training and induces a continuous trust region.

3 Soft Sequence Policy Optimization

While Soft Adaptive Policy Optimization (SAPO) preserves token-level adaptivity through smooth gating, it is sequence-coherent only under two mild assumptions: (i) small policy updates, i.e., $\rho_{i,t}(\theta) \approx 1$, and (ii) low intra-sequence dispersion of importance ratios, $\frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\big(\log \rho_{i,t}(\theta) - \log s_i(\theta)\big)^2$. These assumptions may be violated in practice, particularly for long sequences or during the early stages of training. We therefore aim to redesign the SAPO objective to relax these assumptions while retaining token-level adaptivity.

In the absence of clipping, the geometric mean of token-wise importance ratios coincides with the length-normalized sequence-level importance ratio used in GSPO. GMPO can therefore be interpreted as a relaxation of sequence-level policy optimization that preserves token-level adaptivity.

Building on this observation, we propose *Soft Sequence Policy Optimization* (SSPO). SSPO aggregates token-level gating functions using a geometric mean, yielding the objective

$$\mathcal{J}_{\mathrm{SSPO}}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\,\{y_i\}_{i=1}^{G}\sim\pi_{\theta_{\mathrm{old}}}(\cdot|x)} \left[\frac{1}{G}\sum_{i=1}^{G}\left(\prod_{t=1}^{|y_i|} f\Big(\rho_{i,t}(\theta);\,\widehat{A}_i\Big)\right)^{\frac{1}{|y_i|}}\cdot\widehat{A}_i\right]. \quad (11)$$

Here, $f(\,\cdot\,;\widehat{A})$ is a non-negative gating function whose properties are specified below.

Gradient Analysis To analyze the optimization behavior of SSPO, we compare its gradient to that of SAPO.

Differentiating the SAPO objective in Eq. (8) with respect to the model parameters θ yields the weighted policy gradient

$$\nabla_\theta \mathcal{J}_{\mathrm{SAPO}}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\,\{y_i\}_{i=1}^{G}\sim\pi_{\theta_{\mathrm{old}}}(\cdot|x)} \left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y_i|}\sum_{t=1}^{|y_i|} f'_{\mathrm{Soft}}\Big(\rho_{i,t}(\theta);\widehat{A}_{i,t}\Big)\,\rho_{i,t}(\theta)\cdot \cdot\nabla_\theta \log \pi_\theta(y_{i,t}\mid x, y_{i,<t})\,\widehat{A}_{i,t}\right]. \quad (12)$$

For the sigmoid-shaped gate in SAPO,

$$f'_{\mathrm{Soft}}(\rho;\widehat{A}) = 4\,g(\rho;\widehat{A})\big(1-g(\rho;\widehat{A})\big), \qquad g(\rho;\widehat{A}) = \sigma\Big(\tau(\widehat{A})(\rho-1)\Big), \quad (13)$$

the derivative peaks at $\rho = 1$ and exponentially decays as ρ deviates from unity, inducing a soft trust region while preserving on-policy behavior.

Differentiating the SSPO objective in Eq. (11) yields

$$\nabla_\theta \mathcal{J}_{\mathrm{SSPO}}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\,\{y_i\}_{i=1}^{G}\sim\pi_{\theta_{\mathrm{old}}}(\cdot|x)} \left[\frac{1}{G}\sum_{i=1}^{G}\left(\prod_{t=1}^{|y_i|} f(\rho_{i,t}(\theta);\widehat{A}_i)\right)^{\frac{1}{|y_i|}}\cdot\widehat{A}_i\cdot \cdot\frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\frac{f'(\rho_{i,t}(\theta);\widehat{A}_i)}{f(\rho_{i,t}(\theta);\widehat{A}_i)}\,\rho_{i,t}(\theta) \nabla_\theta \log \pi_\theta(y_{i,t}\mid x, y_{i,<t})\right]. \quad (14)$$

Thus, each token-level policy gradient is modulated by two factors: a sequence-level geometric aggregation of token gates and a local soft importance weight defined as $w(\rho;\widehat{A}) \triangleq \frac{f'(\rho;\widehat{A})}{f(\rho;\widehat{A})}\,\rho$.

Gate Design Motivated by Eq. (14) and the Scopic objective (Chen et al., 2023), we require the gating function family $f(\rho;\widehat{A}) : \mathbb{R}_{++} \to \mathbb{R}_+$ to satisfy the following conditions: (i) f is continuously differentiable; (ii) $f(1;\widehat{A}) = 1$ and $f'(1;\widehat{A}) = 1$; (iii) the local weight multiplier forms a bell-shaped curve centered at $\rho = 1$ to suppress outlier importance ratios; (iv) f is monotonically non-decreasing.

Condition (ii) ensures that when $\rho_{i,t}(\theta) = 1$, the gradient contribution coincides with the standard on-policy update. Conditions (ii) and (iv) ensure that the gating function preserves the ordinal relationship of the importance ratios, maintaining consistency with sequence-level importance weighting.

Condition (iii) enforces controlled attenuation for samples deviating from $\rho = 1$, stabilizing training by discouraging aggressive updates from off-policy samples. Furthermore, requiring $w(\rho; \widehat{A}) \to 0$ as $\rho \to \infty$ guarantees that samples with extreme importance ratios contribute negligibly to the gradient, effectively mitigating instability caused by heavy-tailed ratio distributions.

We explore two formalizations of condition (iii). The first (iii.1) requires the weight multiplier $\frac{f'}{f}$ to attain a global maximum at $\rho = 1$ and decrease monotonically as ρ deviates from 1, with $\frac{f'}{f} \to 0$ as $\rho \to \infty$. The second (iii.2) imposes the same properties on the full token-wise weight $w(\rho; \widehat{A}) = \frac{f'(\rho;\widehat{A})}{f(\rho;\widehat{A})} \cdot \rho$, adding $w(0; \widehat{A}) = 0$. We discuss the trade-offs between these formulations in Subsection 4.1.

Based on empirical results provided further, we instantiate the gate as

$$f_{\text{SSPO}}(\rho; \widehat{A}) = \exp\left(\tau(\widehat{A}) \cdot \arctan\left(\frac{\log \rho}{\tau(\widehat{A})}\right)\right), \tag{15}$$

where $\tau(\widehat{A})$ is an advantage-dependent temperature. This choice yields a local weight:

$$w(\rho; \widehat{A}) = \frac{1}{1 + \left(\frac{\log \rho}{\tau(\widehat{A})}\right)^2}, \tag{16}$$

which peaks at unity when $\rho = 1$ and decays for large deviations. Consequently, SSPO yields bounded gradients without hard clipping while preserving unbiased on-policy updates. Specifically, the induced weight $w(\rho; \widehat{A})$ urges a Cauchy-shaped attenuation in log-ratio space. Its heavy tails preserve contributions from rather off-policy tokens, while geometric aggregation and a gentle slope simultaneously suppress outliers in sequence-level product.

Other Design Choices A central design challenge in SSPO is balancing update stability and policy expressiveness. Following SAPO, we employ distinct temperatures for positive and negative advantages. To cause the gradients associated with negative advantage tokens to decay more rapidly, $\tau_{\text{neg}} \geqslant \tau_{\text{pos}}$ is set in accordance with the proposed gate behavior. First, DAPO (Yu et al., 2025) demonstrates that relaxing the effective upper bound on policy updates (Clip-Higher) improves exploration and mitigates entropy collapse. Second, negative-advantage tokens are empirically more destabilizing (Gao et al., 2025): their gradients redistribute probability mass toward many unsampled and often irrelevant tokens, whereas positive advantages primarily sharpen the sampled token's logit.

The final SSPO objective is

$$\mathcal{J}_{\text{SSPO}}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\{y_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|x)} \left[\frac{1}{G}\sum_{i=1}^{G}\left(\prod_{t=1}^{|y_i|} f_{\text{SSPO}}\left(\rho_{i,t}(\theta); \widehat{A}_i\right)\right)^{\frac{1}{|y_i|}} \cdot \widehat{A}_i\right]. \tag{17}$$

4 Experiments

4.1 Gate options

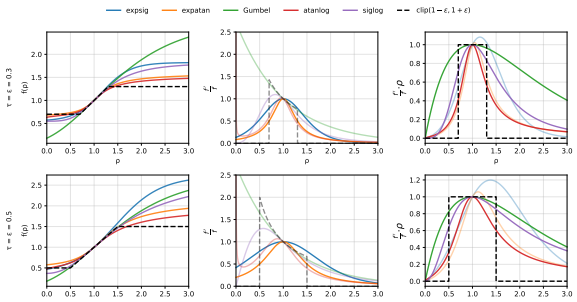


Figure 1: Illustration of the candidate SSPO gate families and their induced weighting behavior as the importance ratio departs from the on-policy point $\rho = 1$. The heavier-tailed arctan-based gates preserve more signal away from the trust region, while sigmoid-style gates attenuate more aggressively.

Let $\tau_{\widehat{A}} := \tau(\widehat{A})$ for brevity. We evaluate several candidate gating functions $f(\rho; \widehat{A})$ that satisfy requirements described above:

$$f_{\text{expsig}}(\rho; \widehat{A}) = \exp\big(4\tau_{\widehat{A}} \cdot \sigma(\tfrac{\rho-1}{\tau_{\widehat{A}}}) - 2\tau_{\widehat{A}}\big), \tag{18}$$

$$f_{\text{expatan}}(\rho; \widehat{A}) = \exp\big(\tau_{\widehat{A}} \cdot \arctan(\tfrac{\rho-1}{\tau_{\widehat{A}}})\big), \tag{19}$$

$$f_{\text{Gumbel}}(\rho) = \exp\big(-\exp(-(\rho-1)) + 1\big), \tag{20}$$

$$f_{\text{atanlog}}(\rho; \widehat{A}) = \exp\big(\tau_{\widehat{A}} \cdot \arctan(\tfrac{\log \rho}{\tau_{\widehat{A}}})\big), \tag{21}$$

$$f_{\text{siglog}}(\rho; \widehat{A}) = \exp\big(4\tau_{\widehat{A}} \cdot \sigma(\tfrac{\log \rho}{\tau_{\widehat{A}}}) - 2\tau_{\widehat{A}}\big). \tag{22}$$

We group these gates by the formalization of the condition (iii) they satisfy. Gates f_{expsig} and f_{expatan} enforce unimodality on the ratio-level multiplier $\frac{f'(\rho;\widehat{A})}{f(\rho;\widehat{A})}$ (condition (iii.1)). In contrast, f_{Gumbel}, f_{atanlog}, and f_{siglog} enforce unimodality on the full token-wise weight $w(\rho; \widehat{A})$ (condition (iii.2)). Figure 1 illustrates gate behavior in sequence-level aggregation and token-wise weights compared to hard clipping. Across these variants,

the practical distinction is between more aggressive sigmoid-style attenuation and heavier-tailed $\arctan$ attenuation.

The gate f_{Gumbel} is parameter-free, as it does not allow temperature parameterization while simultaneously satisfying condition (ii) and the $\arg\max$ restriction of (iii). Empirically, absence of hyperparameters makes it a strong baseline among SSPO gates.

4.2 Experimental Setup

SSPO is compared against GRPO, SAPO, and two sequence-level objectives: GMPO in the ablation study and GSPO in the transfer study.

Reward is computed as a sum of the formatting reward and indicator of the answer correctness. More details are available in Appendix A.

Hyperparameters We use group size $G = 8$ responses per prompt and the maximum completion length of 512 tokens. We do not apply the KL regularization term and update the rollout policy every 2 steps. Clipping hyperparameters, such as ε, τ_{neg}, and τ_{pos}, are selected via grid search. See Appendix A for more details. Unless noted otherwise, we report results for the best configuration per method. For the transfer ablation below, we study a harsher stale-policy regime with the rollout policy being refreshed every 4 optimizer steps.

Evaluation protocol For evaluation, we use zero-shot prompting and greedy decoding for all models. We extract and verify the final answer using the same procedure as in training.

4.3 Ablation Study

We fine-tune Qwen2.5-7B-Instruct from a cold-start initialization. We focus on mathematical reasoning tasks with verifiable outcome-level rewards. The training dataset comprises problems from GSM8K (Cobbe et al., 2021) and DeepMath-103K (He et al., 2025). We report Pass@1 on GSM8K test set, MATH-500 (Lightman et al., 2023), and AIME 2025 in Table 1. Training curves may be found at the Appendix B.

Comparison to SAPO and GMPO helps to isolate the effects of geometric-mean aggregation and soft clipping. SSPO with the f_{expsig} and f_{siglog} gates are the closest counterparts to SAPO, as the gates share the same attenuation regime.

Across all objectives, fine-tuning improves over the base Qwen2.5-7B-Instruct model. Among the baselines, GMPO is the strongest overall, reaching

Model	GSM8K	M500	A25	Δ Avg
Qwen2.5-7B-Instruct	77.9	56.6	10.0	–
+ GRPO	84.5	64.2	10.0	+4.7
+ GMPO	*88.8*	*65.0*	*13.3*	*+7.5*
+ SAPO	82.8	58.6	10.0	+2.3
+ SSPO (f_{expsig})	89.2	64.8	<u>16.7</u>	+8.7
+ SSPO (f_{expatan})	<u>89.6</u>	<u>65.8</u>	<u>16.7</u>	+9.2
+ SSPO (f_{Gumbel})	88.9	63.8	<u>16.7</u>	+8.3
+ SSPO (f_{siglog})	88.9	64.8	<u>16.7</u>	+8.8
+ SSPO (f_{atanlog})	**91.6**	**66.6**	**20.0**	**+11.2**

Table 1: Pass@1 on mathematical reasoning benchmarks. The best results are shown in bold, the second-best are underlined and the strongest baseline in italics. Δ Avg reports the average absolute improvement over the base model across the three benchmarks. A25 denotes AIME25; M500 denotes MATH-500.

88.8 on GSM8K, 65.0 on MATH-500, and 13.3 on AIME 2025. SSPO improves upon GMPO on all three benchmarks; the best-performing instantiation, SSPO with the f_{atanlog} gate, reaches 91.6 on GSM8K, 66.6 on MATH-500, and 20.0 on AIME 2025. The largest relative gain appears on AIME 2025, where SSPO increases Pass@1 from 4/30 to 6/30 solved problems, consistent with the hypothesis that sequence-level geometric aggregation with token-level soft weighting is especially helpful when long, high-variance trajectories make off-policy updates unstable. These results in math reasoning are the empirical anchor for the following transfer study.

4.4 Cross-Model Transfer Study

Model	Method	A24	A25	M500	CodeAct
Qwen3-8B	Base	17.8	13.2	56.8	18.4
	GRPO	23.9	18.7	61.2	19.3
	GSPO	26.1	20.8	63.8	24.0
	SAPO	25.4	21.5	62.9	22.4
	SSPO	**28.3**	**22.6**	**64.6**	**27.0**
Qwen3-14B	Base	21.4	15.8	60.7	21.8
	GRPO	29.6	22.8	66.5	26.5
	GSPO	32.8	25.4	**69.8**	**30.4**
	SAPO	31.4	26.3	68.9	28.6
	SSPO	**35.1**	**27.9**	69.5	29.6
Qwen3-30B-A3B	Base	24.8	18.5	63.1	23.6
	GRPO	33.7	25.0	69.4	17.0
	GSPO	37.9	28.1	73.6	28.9
	SAPO	36.8	**29.4**	73.1	24.0
	SSPO	**40.6**	29.0	**74.2**	**32.9**

Table 2: Benchmark matrix for the transfer settings. A24/A25 denote AIME24/AIME25 Pass@1; M500 denotes MATH-500 Pass@1; CodeAct denotes end-to-end solved rate.

To assess whether the observed on 7B parameters model advantage remains under larger models, mixture-of-experts routing, and other domains, a transfer study is conducted. Qwen3-8B,

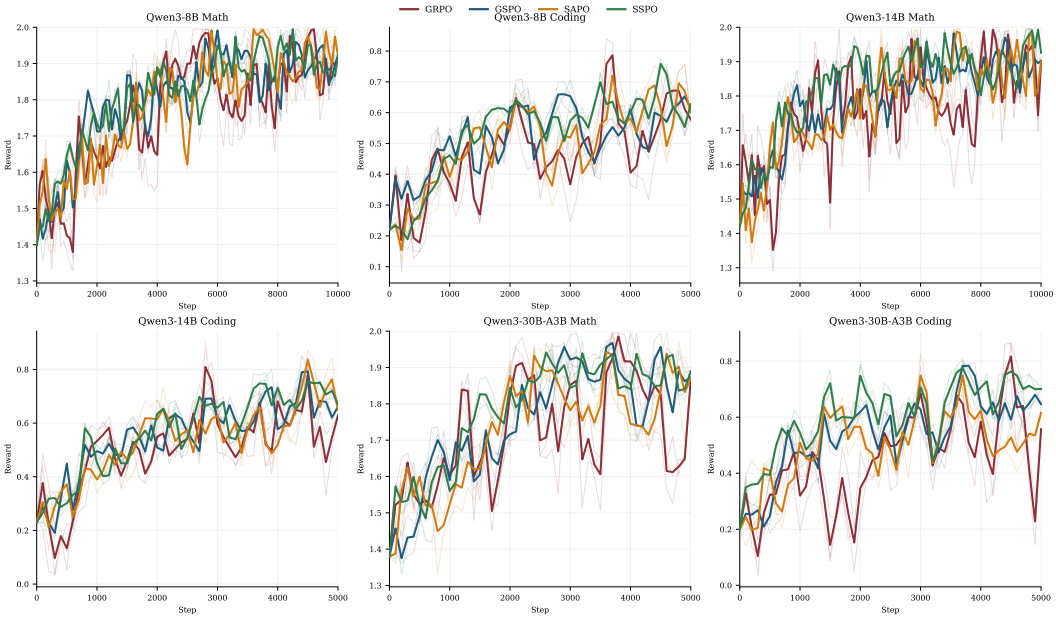


Figure 2: Sequence reward curves for all six $\mu = 4$ transfer settings arranged as a 2×3 grid: Qwen3-8B, Qwen3-14B, and Qwen3-30B-A3B on math and coding. Thick curves denote method means, while thin curves show the three per-setting seeds without exponential smoothing. On the math tasks, reward stays within the natural $[0, 2]$ range induced by answer and format scoring.

Qwen3-14B, and Qwen3-30B-A3B are fine-tuned on math and coding. The math settings assume DAPO-103K training and evaluation on AIME24, AIME25, and MATH-500; the coding settings assume CodeAct (Wang et al., 2024) training and CodeAct test evaluation. Table 2 groups the results by backbone and includes the corresponding public baseline for each model.

Several patterns recur across all the six settings. First, all four objectives improve over the base models, but the gains vary across benchmarks. SSPO delivers the strongest overall progression in math on the dense 8B and 14B backbones, leading on both AIME sets, and reaches the best MATH-500 value on Qwen3-8B. Second, GSPO is observed to be the most competitive hard-clipping baseline. It attains the best MATH-500 and CodeAct scores on Qwen3-14B, indicating that sequence-level hard weighting can remain slightly more conservative yet efficient when optimization stays relatively stable. Third, the MoE setting clearly separates methods. On Qwen3-30B-A3B, SSPO produces the strongest AIME24, MATH-500, and CodeAct results. At the same time, SAPO attains the best AIME25 value, suggesting that its softer token-level attenuation can still be useful on the hardest long-horizon slice, while SSPO is stronger overall.

Figures 2 and 3 present the optimization panel, reporting reward and entropy dynamics. Under the regime of increased policy staleness, GRPO exhibits the deepest entropy collapse and the largest reward shocks, GSPO stabilizes the hard sequence-level objective, and SSPO maintains the strongest reward gains in most settings. Accordingly, the benchmark matrix shows the clearest downstream gains of the proposed objective.

5 Related Works

Group-based policy optimization has become a strong baseline for reinforcement learning with large language models, particularly in alignment and reasoning tasks. Methods such as GRPO (Shao et al., 2024) and RLOO (Ahmadian et al., 2024) estimate advantages by sampling multiple responses per prompt and computing relative rewards within each group, avoiding an auxiliary critic network. GRPO further enables off-policy updates leveraging importance-weighted updates, making it a practical default in large-scale RL pipelines and inducing various modifications (Yu et al., 2025; Liu et al., 2025b,a).

In realistic training setups, off-policy learning is unavoidable due to large rollout buffers and mini-batch gradient steps. However, applying token-level importance sampling ratios to sequence-level rewards introduces a policy-reward mismatch that

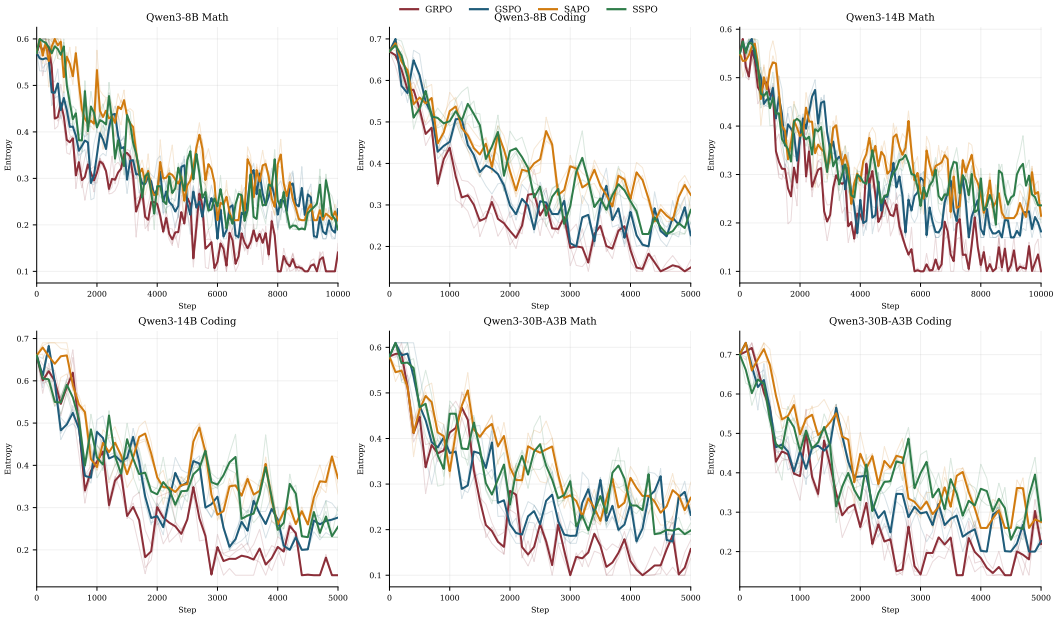


Figure 3: Token entropy curves for the same six $\mu = 4$ transfer settings, again arranged as a 2×3 grid. The entropy panels make the contrast in collapse behavior visible: GRPO shows the sharpest entropy loss, GSPO stabilizes hard sequence-level clipping, and SSPO preserves high-reward exploration in most settings without flattening the trajectories into unrealistically smooth traces.

can destabilize training. Recent work addresses this issue by enforcing sequence-level coherence in policy updates. GSPO (Zheng et al., 2025) and GMPO (Zhao et al., 2025) substitute arithmetic mean of token-level ratios with sequence-consistent formulations, improving stability and variance control.

Hard clipping of importance ratios, introduced in PPO (Schulman et al., 2017), remains a common strategy for variance reduction in off-policy optimization (Shao et al., 2024; MiniMax et al., 2025). While effective, clipping induces a fundamental bias-variance tradeoff: aggressive clipping improves stability but harms sample efficiency and exploration, whereas loose clipping yields noisy updates (Chen et al., 2023; Dwyer et al., 2025). Several works explore alternatives, including asymmetric clipping (Yu et al., 2025) and gradient-preserving approaches (Su et al., 2025a), but these methods remain sensitive to hyperparameters.

An alternative line of work advocates for soft, entropy-aware objectives that avoid hard clipping. SAPO (Gao et al., 2025) and related soft trust-region methods (Dwyer et al., 2025) demonstrate improved robustness in off-policy settings by smoothly attenuating large importance ratios. Recent work also explores global soft constraints on policy distributions, such as Entropy Ratio Clipping (Su et al., 2025b), which uses the ratio of policy entropies between updates to constrain distributional drift beyond standard importance clipping. However, these approaches are primarily developed for token-level objectives and do not explicitly address sequence-level credit assignment in group-based learning.

6 Conclusion

The paper introduces Soft Sequence Policy Optimization (SSPO), a sequence-level objective for off-policy RL fine-tuning of LLMs. The design of the proposed objective combines alignment between the reward unit and the importance correction unit with a flexible token-wise gate function. We analyze the resulting gradients and derive practical gate design criteria. Empirically, SSPO improves training stability and outperforms strong group-based baselines on mathematical reasoning, while the cross-model transfer study suggests that the same mechanism remains competitive across larger dense models, MoE backbones, and coding tasks. While GSPO and SAPO stay competitive, SSPO attains the broadest overall advantage across the combined math and coding setups. Overall, our findings suggest that combining sequence-level geometric aggregation with heavy-tailed bounded token gating provides a favorable stability-efficiency trade-off for off-policy sequence optimization.

Limitations

SSPO assumes sequence-level rewards over full responses and is not directly tailored to fine-grained token-level reward models. Our measured experiments cover only one verifiable math setup, while the cross-model ablation is a calibrated transfer study rather than completed training on all settings. Finally, SSPO introduces gate and temperature choices that still require tuning, so broader sensitivity analysis remains future work.

Computational Resources

The research was carried out using the MSU-270 supercomputer of Lomonosov Moscow State University.

Acknowledgements

This work was supported by the The Ministry of Economic Development of the Russian Federation in accordance with the subsidy agreement (agreement identifier 000000C313925P4H0002; grant No 139-15-2025-012).

References

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

Xing Chen, Dongcui Diao, Hechang Chen, Hengshuai Yao, Haiyin Piao, Zhixiao Sun, Zhiwei Yang, Randy Goebel, Bei Jiang, and Yi Chang. 2023. The sufficiency of off-policyness and soft clipping: Ppo is still insufficient according to an off-policy measure. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 7078–7086.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Madeleine Dwyer, Adam Sobey, and Adriane Chapman. 2025. It's not you, it's clipping: A soft trust-region via probability smoothing for llm rl. *Preprint*, arXiv:2509.21282.

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

Zhiwei He, Tian Liang, Jiahao Xu, Qiuzhi Liu, Xingyu Chen, Yue Wang, Linfeng Song, Dian Yu, Zhenwen Liang, Wenxuan Wang, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. 2025. Deepmath-103k: A large-scale, challenging, decontaminated, and verifiable mathematical dataset for advancing reasoning. *Preprint*, arXiv:2504.11456.

Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron Courville, and Nicolas Le Roux. 2025. Vineppo: Refining credit assignment in rl training of llms. *Preprint*, arXiv:2410.01679.

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

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

Zihe Liu, Jiashun Liu, Yancheng He, Weixun Wang, Jiaheng Liu, Ling Pan, Xinyu Hu, Shaopan Xiong, Ju Huang, Jian Hu, Shengyi Huang, Johan Obando-Ceron, Siran Yang, Jiamang Wang, Wenbo Su, and Bo Zheng. 2025b. Part i: Tricks or traps? a deep dive into rl for llm reasoning. *Preprint*, arXiv:2508.08221.

MiniMax, :, Aili Chen, Aonian Li, Bangwei Gong, Binyang Jiang, Bo Fei, Bo Yang, Boji Shan, Changqing Yu, Chao Wang, Cheng Zhu, Chengjun Xiao, Chengyu Du, Chi Zhang, Chu Qiao, Chunhao Zhang, Chunhui Du, Congchao Guo, and 109 others. 2025. Minimax-m1: Scaling test-time compute efficiently with lightning attention. *Preprint*, arXiv:2506.13585.

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

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

Zhenpeng Su, Leiyu Pan, Minxuan Lv, Tiehua Mei, Zijia Lin, Yuntao Li, Wenping Hu, Ruiming Tang, Kun Gai, and Guorui Zhou. 2025b. Entropy ratio clipping as a soft global constraint for stable reinforcement learning. *Preprint*, arXiv:2512.05591.

Xingyao Wang, Yangyi Chen, Lifan Yuan, Yizhe Zhang, Yunzhu Li, Hao Peng, and Heng Ji. 2024. Executable code actions elicit better llm agents. In *Forty-first International Conference on Machine Learning*.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, and 16 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *Preprint*, arXiv:2503.14476.

Yuzhong Zhao, Yue Liu, Junpeng Liu, Jingye Chen, Xun Wu, Yaru Hao, Tengchao Lv, Shaohan Huang, Lei Cui, Qixiang Ye, Fang Wan, and Furu Wei. 2025. Geometric-mean policy optimization. *Preprint*, arXiv:2507.20673.

Chujie Zheng, Pei Ke, Zheng Zhang, and Minlie Huang. 2023. Click: Controllable text generation with sequence likelihood contrastive learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1022–1040.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. 2025. Group sequence policy optimization. *Preprint*, arXiv:2507.18071.

A Experimental Setup Details

Reward and answer verification As for the reward, we use $r = r_{\text{answer}} + r_{\text{format}}$, where $r_{\text{answer}} = \mathbb{I}[\hat{a} = a]$ and $\hat{a}$ is extracted from the response between special tokens `<answer>` and `</answer>`. The formatting reward encourages the template `<think>` $\cdots$ `</think>` `<answer>` $\cdots$ `</answer>`. $r_{\text{format}} \in \{1, 0.5, 0.25, 0\}$ with scores assigned as fully correct format/ correct edge tokens (both) / correct edge (either) / otherwise.

Hyperparamater tuning For the main experimental series, clipping hyperparameters are selected via search over the following grid: ε for GRPO and GMPO in $\{0.1, 0.2, 0.3\}$; $(\tau_{\text{neg}}, \tau_{\text{pos}})$ for SAPO in $\{(1, 1.05), (3, 4), (7, 8)\}$; and $(\tau_{\text{neg}}, \tau_{\text{pos}})$ for each SSPO gate in $\{(0.3, 0.2), (0.6, 0.5)\}$.

Training Datasets Description For training the following datasets are used: GSM8K (Cobbe et al., 2021), DeepMath-103K (He et al., 2025), and CodeActInstruct (Wang et al., 2024). GSM8K contains curated grade-school math problems requiring multi-step reasoning. DeepMath-103K spans a diverse spectrum of mathematical subjects, including algebra, calculus, number theory, geometry, probability, and discrete mathematics, providing problems of widely varying difficulty. CodeActInstruct is an instruction-tuning dataset for training large language model agents to act through executable Python code rather than fixed-form tool calls. The dataset was introduced with the CodeAct framework to improve agents' ability to perform flexible tool use, while preserving general language-model capabilities when mixed with standard instruction-tuning data.

B More on Ablation Study

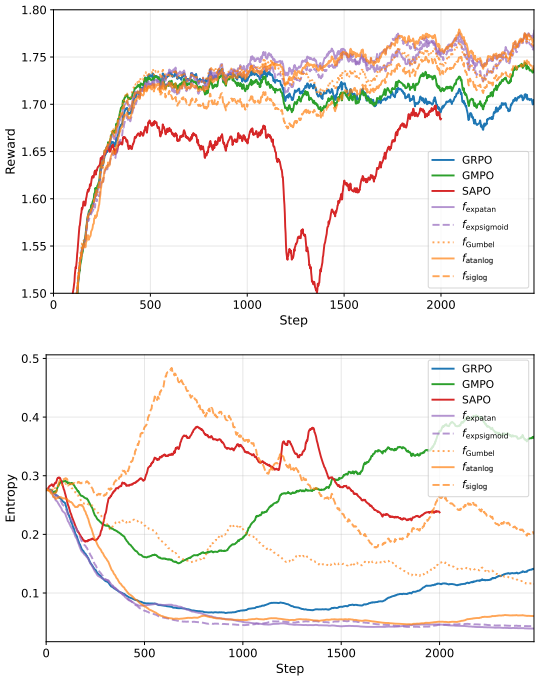


Figure 4: Training curves for reward and entropy. Values are reported after exponential smoothing with $\alpha = 0.02$ for best configurations of each objective selected by grid search.

Figure 4 shows the smoothed training reward and token entropy for the best hyperparameter setting of each method. We observe that SAPO may suffer from reward collapse in our setting, whereas SSPO maintains stable learning dynamics and achieves consistently higher rewards throughout training.

Gate choice introduces two largely independent design axes. The first is defined by bell-shaped multiplier: on the ratio-level multiplier $\frac{f'}{f}$ (formalization (iii.1)) versus on the full token-wise weight $w(\rho) = \frac{f'}{f}\rho$ with $w(0) = 0$ (formalization (iii.2)). The second is the attenuation profile: sigmoid-based gates suppress off-policy tokens with expo-

nentially decaying weights as the (log-)ratio deviates from 1, whereas $\arctan$-based gates induce a Cauchy-shaped attenuation in log-ratio space with heavier tails.

Empirically, $\arctan$-based gates are more robust than sigmoid-based gates across benchmarks. We hypothesize that this behavior reflects a better trade-off: the sequence-level geometric aggregation in SSPO already reduces sensitivity to extreme outliers, so a heavier-tailed token gate can retain useful learning signal from quite off-policy tokens without destabilizing updates. In contrast, aggressively down-weighting such tokens can reduce effective sample reuse and slow down improvement on harder problems.