MASPO: Unifying Gradient Utilization, Probability Mass, and Signal Reliability for Robust and Sample-Efficient LLM Reasoning

Xiaoliang Fu[1,2,*], Jiaye Lin[1,*,†], Yangyi Fang[1,3,*], Binbin Zheng[1,4], Chaowen Hu[1], Zekai Shao[2], Cong Qin[1,5], Lu Pan[1], Ke Zeng[1], Xunliang Cai[1]
[1]Meituan [2]Fudan University [3]Tsinghua University
[4]University of Science and Technology of China [5]Peking University
{fuxiaoliang04, linjiaye}@meituan.com

Abstract

Existing Reinforcement Learning with Verifiable Rewards (RLVR) algorithms, such as GRPO, rely on rigid, uniform, and symmetric trust region mechanisms that are fundamentally misaligned with the complex optimization dynamics of Large Language Models (LLMs). In this paper, we identify three critical challenges in these methods: (1) *inefficient gradient utilization* caused by the binary cutoff of hard clipping, (2) *insensitive probability mass* arising from uniform ratio constraints that ignore the token distribution, and (3) *asymmetric signal reliability* stemming from the disparate credit assignment ambiguity between positive and negative samples. To bridge these gaps, we propose **M**ass-**A**daptive **S**oft **P**olicy **O**ptimization (**MASPO**), a unified framework designed to harmonize these three dimensions. MASPO integrates a differentiable soft Gaussian gating to maximize gradient utility, a mass-adaptive limiter to balance exploration across the probability spectrum, and an asymmetric risk controller to align update magnitudes with signal confidence. Extensive evaluations demonstrate that MASPO serves as a robust, all-in-one RLVR solution, significantly outperforming baselines. Our code is available at: https://github.com/FlyTune/MASPO-RL.

1 Introduction

Reinforcement Learning with Verifiable Rewards (RLVR) has evolved into the cornerstone of reasoning optimization in Large Language Models (LLMs) (Lightman et al., 2023b; Shao et al., 2024). While algorithms like GRPO (Shao et al., 2024) eliminate the need for specific critic models, they rely on a "hard clipping" mechanism inherited from PPO (Schulman et al., 2017). We rethink that it is a fragmented solution, failing to address the complex interplay between heavy-tailed vocabulary distributions and sparse reward signals.

As illustrated in Figure 1, current paradigms suffer from three distinct misalignments: (1) *Inefficient Gradient Utilization*. Hard clipping mechanism imposes a rigid binary cutoff that indiscriminately discards valuable reinforcement signals from successful explorations exceeding the predefined boundary, while simultaneously failing to provide corrective gradients for severe errors that are within the boundary, thus diminishing the effective utilization of generated samples. (2) *Insensitive Probability Mass*. A uniform clipping range (e.g., $[1-\varepsilon, 1+\varepsilon]$) neglects the token's absolute probability. It imposes overly loose constraints on high-probability tokens (risking policy collapse) while excessively restricting low-probability tokens in the long tail (hindering exploration). (3) *Asymmetric Signal Reliability*. Positive rewards verify correct reasoning, whereas negative rewards suffer from credit assignment ambiguity. Treating their advantage estimates symmetrically overlooks their disparate Signal-to-Noise Ratios (SNR).

Aiming to address these challenges holistically, we propose **MASPO**, a **M**ass-**A**daptive **S**oft **P**olicy **O**ptimization framework, which designed to synthesize trust region management into a single, cohesive objective. By replacing the rigid box constraint with a soft Gaussian gating mechanism, MASPO resolves the gradient inefficiency. Crucially, it synergizes two adaptive components: a mass-adaptive limiter that expands the exploration budget for low-probability tokens to address mass insensitivity and an asymmetric risk controller that modulates update magnitudes based on signal confidence to resolve reliability asymmetry. Our contributions in this paper are summarized as follows:

- **Unified Perspective:** We systematically identify inherent challenges in current trust region paradigms and propose a holistic perspective that aligns RLVR optimization by addressing three fundamental misalignments.

[*] Equal contribution. [†] Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 41348–41365
July 2-7, 2026

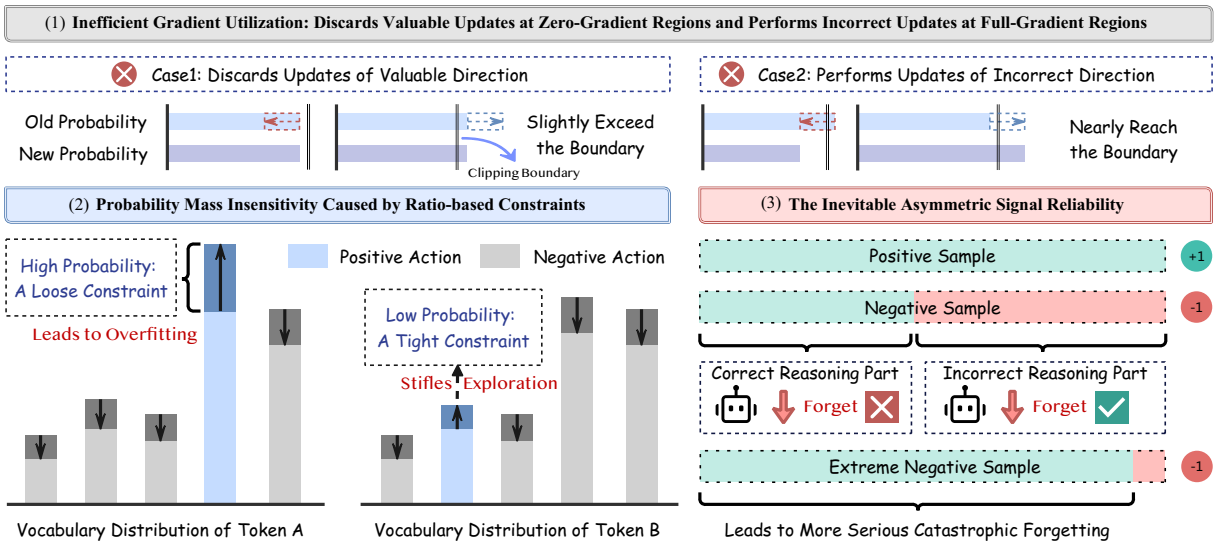


Figure 1: MASPO resolves three core limitations of GRPO: (1) *inefficient gradient utilization* caused by hard boundaries that indiscriminately discard valid gradients; (2) *insensitive probability mass*, where uniform clipping overlooks the head/tail mass disparity; and (3) *asymmetric signal reliability*, which neglects the inherent noise differences between verified positive and ambiguous negative signals.

- **All-in-One Framework:** We propose a comprehensive solution, which unifies a soft Gaussian gating, a mass-adaptive limiter, and an asymmetric risk controller into a single framework.
- **Superior Performance:** Comprehensive evaluations on diverse benchmarks demonstrate that MASPO achieves superior sample efficiency and reasoning performance. Further experiments confirm its robustness and effectiveness in stabilizing long-chain reasoning.

2 Related Works

2.1 RLVR in LLMs

RLVR enhances LLM reasoning using rule-based signals, unlike preference-based RLHF (Ouyang et al., 2022). While PPO (Schulman et al., 2017) is a standard approach, its critic dependence limits scalability. Recent work favors outcome-based supervision (Uesato et al., 2022) via group-based optimization. GRPO (Shao et al., 2024) eliminates the critic by normalizing rewards within groups. However, it struggle with sparse rewards and high variance, necessitating stable variants (Bai et al., 2026). Beyond optimization, recent work also seeks to enhance reasoning quality through improved model architectures (Dong et al., 2026) and inference-time strategies (Jiang et al., 2026).

2.2 Optimization Directions for GRPO

Recent studies have attempted to address the limitations of the standard GRPO framework. We categorize these efforts according to the three critical misalignments identified in this paper:

Addressing Inefficient Gradient Utilization. Standard hard clipping discards valuable gradients from exploratory tokens that exceed the trust region boundary. To mitigate this *Inefficient Gradient Utilization*, soft clipping strategies have been proposed to retain partial gradients. Approaches like CISPO (Chen et al., 2025a), GPPO (Su et al., 2025a), CE-GPPO (Su et al., 2025b) and ASPO (Wang et al., 2025) introduce smooth piecewise decay functions. SAPO (Gao et al., 2025) goes a step further by removing boundaries entirely, employing a global gating mechanism.

Addressing Insensitive Probability Mass. Uniform Importance Sampling (IS) (Precup et al., 2000) constraints overly restrict exploration in low-probability regions, reflecting the issue of *insensitive probability mass*. While entropy regularization (Williams and Peng, 1991; O'Donoghue et al., 2016; Eysenbach and Levine, 2021) encourages distribution uniformity, it indiscriminately encourages all (He et al., 2025; Huang et al., 2025). Clip Higher (Yu et al., 2025) relaxes upper bounds globally but ignores token-specific probabilities. DAC (Yang et al., 2025a), a more aligned method, adapts bounds based on policy probabilities.

Addressing Asymmetric Signal Reliability. Positive samples are crucial for LLM reasoning but are frequently outweighed by negative ones in

GRPO (Xiong et al., 2025), highlighting the issue of *Asymmetric Signal Reliability*. Methods like Advantage Reweighting (Yang et al., 2025b) and BAPO (Xi et al., 2025) boost positive sample impact by modifying advantages or clipping bounds. Other approaches achieve implicit rebalancing via relaxed bounds (Yu et al., 2025), asymmetric soft clipping (Su et al., 2025b), or adding entropy terms to the advantage (Cheng et al., 2025).

3 Preliminary

Problem Formulation. We consider the task of optimizing an LLM for reasoning tasks within the framework of Reinforcement Learning with Verifiable Rewards (RLVR). Let $\mathcal{D} = \{q\}$ represent a dataset of queries, where the policy π_θ generates a response o for a given input q. The correctness of the generated response is evaluated by a deterministic, rule-based reward function $r(q, o)$. Adopting the Group Relative Policy Optimization (GRPO) (Shao et al., 2024) paradigm, for each query q, we sample a group of G outputs $\{o_1, o_2, \ldots, o_G\}$ from the current old policy $\pi_{\theta_{\text{old}}}$. The advantage estimate $\hat{A}_i$ for the i-th output is derived by standardizing the rewards within the group: $\hat{A}_i = (r_i - \mu_r)/\sigma_r$, where μ_r and σ_r denote the mean and standard deviation of the rewards within the sampled group, respectively.

Gradient Estimation and Utilization Efficiency. To analyze the *Inefficient Gradient Utilization* caused by rigid constraints, we formulate the policy gradient (Williams, 1992) in a generalized form. Let $\rho_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|q,o_{i,<t})}$ be the importance sampling ratio. The objective can be unified as:

$$\nabla_\theta \mathcal{J}(\theta) = \mathbb{E}_{q\sim\mathcal{D},\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q)} \frac{1}{\sum_{i=1}^{G}|o_i|} \sum_{i=1}^{G} \sum_{t=1}^{|o_i|} \mathcal{F}_{i,t}(\rho)\rho_{i,t}(\theta)\hat{A}_i \nabla_\theta \log \pi_\theta(o_{i,t}|q, o_{i,<t}), \quad (1)$$

where $\mathcal{F}_{i,t}(\rho)$ is the gradient weighting function. Standard methods like GRPO employ a binary weighting (hard clipping) where $\mathcal{F}(\rho) = \mathbb{I}(|\rho - 1| \le \varepsilon)$, which zeros out gradients outside the trust region. Attempts to improve utilization, such as SAPO (Gao et al., 2025), propose a continuous decay function to smooth this boundary:

$$\mathcal{F}_{i,t}^{\text{SAPO}}(\rho) = 4p_{i,t}(\theta)(1 - p_{i,t}(\theta))$$
$$p_{i,t}(\theta) = \sigma\left(\tau_{i,t}(\rho_{i,t}(\theta) - 1)\right), \quad (2)$$

where $\sigma(x) = 1/(1 + e^{-x})$ is the sigmoid function.

Probability Mass Sensitivity in Constraints. To address *insensitive probability mass*, prior works have attempted to adjust constraints based on distribution characteristics. Entropy regularization (O'Donoghue et al., 2016) adds a generic penalty $\beta_{\text{E}}\nabla_\theta\mathcal{H}(\pi_\theta)$. Clip Higher (Yu et al., 2025) proposes decoupled boundaries: ε_{low} and $\varepsilon_{\text{high}}$. DAC (Yang et al., 2025a) provides a more rigorous formulation by defining the boundaries as functions of the old policy probability $\pi_{\theta_{\text{old}}}$, acknowledging that low-probability tokens require wider trust regions:

$$0.5 + \frac{1}{2}\sqrt{\max(1 - \frac{4\varepsilon_{\text{low}}}{\pi_{\theta_{\text{old}}}}, 0)} \le \rho_{i,t}(\theta) \le 0.5 + \frac{1}{2}\sqrt{1 + \frac{4\varepsilon_{\text{high}}}{\pi_{\theta_{\text{old}}}}}. \quad (3)$$

While theoretically grounded, this approach retains the rigid "box constraint" structure rather than a continuous adaptation.

Asymmetric Signal Handling. To address *Asymmetric Signal Reliability*, several methods skew contributions toward positive samples. Advantage Reweighting (Yang et al., 2025b) statistically increases positive contributions: $\hat{A}_{i,t} = [\alpha_{\text{A}} \cdot \pi_\theta(o_{i,t}) + (1 - \alpha_{\text{A}})] \cdot \hat{A}_{i,t}^{\text{old}}$. Entropy Advantage (Cheng et al., 2025) adds entropy to the advantage term. BAPO (Xi et al., 2025) explicitly dynamically adjusts the clipping bounds to ensure positive sample contribution volume:

$$\frac{\left|\sum_{\hat{A}} \pi_\theta \left[\min\left(\rho_t\hat{A}, \text{clip}(\rho_t, c_{\text{low}}, c_{\text{high}})\hat{A}\right)\right]\right|}{\left|\sum_{\hat{A}>0} \pi_\theta \left[\min\left(\rho_t\hat{A}, \text{clip}(\rho_t, 0, c_{\text{high}})\hat{A}\right)\right]\right|} \ge \rho_0, \quad (4)$$

where ρ_0 is a threshold parameter, and ρ_t is IS ratio.

4 Methodology

In this section, we present **M**ass-**A**daptive **S**oft **P**olicy **O**ptimization (MASPO). As illustrated in Figure 1, standard RLVR paradigms like GRPO rely on fragmented constraints that create three fundamental misalignments in LLM optimization:

- **Inefficient Gradient Utilization:** Hard clipping imposes a binary cutoff that discards valuable directional gradients from exploratory samples exceeding the boundary, thereby significantly diminishing the effective utilization of

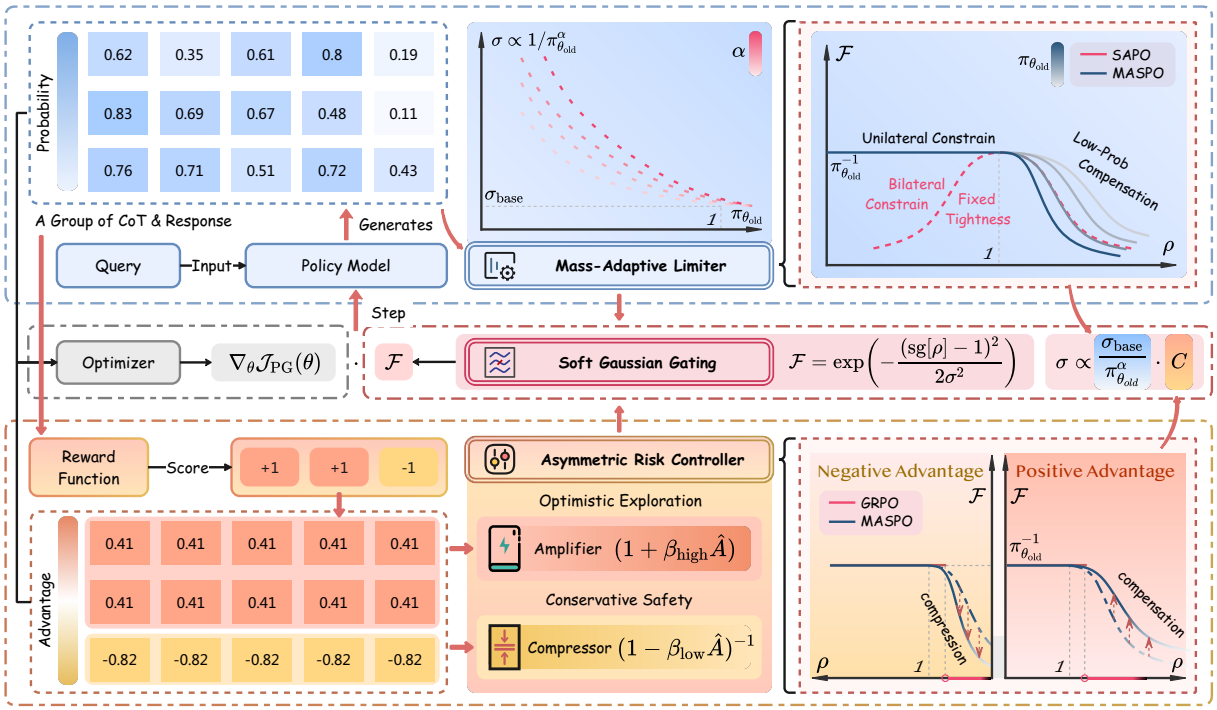


Figure 2: Overview of the **MASPO** framework. The architecture integrates a **Mass-Adaptive Limiter** to scale constraints inversely with token probability and an **Asymmetric Risk Controller** to modulate update magnitude based on advantage signals, unified via a differentiable **Soft Gaussian Gating** mechanism.

informative gradient signals, consequently slowing down the overall convergence speed.

- **Insensitive Probability Mass:** Uniform ratio constraints ignore the vast disparity in token probabilities, failing to account for the massive mass displacement in head tokens versus the negligible shift in tail tokens.
- **Asymmetric Signal Reliability:** Symmetric advantage handling ignores the disparate signal-to-noise ratios between verified positive solutions and ambiguous negative ones.

To address these challenges holistically, MASPO proposes a **unified** framework (see Figure 2). We replace the rigid hard clipping with a **Soft Gaussian Gating** mechanism to ensure gradient continuity (§4.1). Crucially, we regulate this gate via a **Dual-Variable Adaptive Variance** (§4.2), which synergizes a **Mass-Adaptive Limiter** for distribution stability and an **Asymmetric Risk Controller** for signal-aware optimization, ultimately enabling more robust and efficient policy learning.

4.1 Soft Gaussian Gating

To resolve the gradient discontinuity caused by hard clipping, we formulate the MASPO objective using a differentiable confidence score derived from the Principle of Maximum Entropy. The objective is defined as:

$$\mathcal{J}_{\mathrm{MASPO}}(\theta) = \mathbb{E}_{q\sim\mathcal{D},\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\mathrm{old}}}(\cdot|q)} \frac{1}{\sum_{i=1}^{G}|o_i|} \sum_{i=1}^{G}\sum_{t=1}^{|o_i|} \mathcal{F}_{i,t}^{\mathrm{MASPO}} \cdot \rho_{i,t}(\theta)\hat{A}_i, \tag{5}$$

Here, $\mathcal{F}_{i,t}^{\mathrm{MASPO}}$ is the **Soft Gaussian Gating** factor. Unlike hard clipping or SAPO's dual-clip (Ye et al., 2020), this **unilateral design** selectively attenuates aggressive overshoots without hindering conservative updates of lagging tokens:

$$\mathcal{F}_{i,t}^{\mathrm{MASPO}} = \begin{cases} \exp\left(-\frac{(\mathrm{sg}[\rho_{i,t}(\theta)]-1)^2}{2\sigma_{\mathrm{pos}}^2}\right) & \text{if } \hat{A}_{i,t} > 0 \\ & \wedge\ \rho_{i,t}(\theta) > 1 \\ \exp\left(-\frac{(\mathrm{sg}[\rho_{i,t}(\theta)]-1)^2}{2\sigma_{\mathrm{neg}}^2}\right) & \text{if } \hat{A}_{i,t} < 0 \\ & \wedge\ \rho_{i,t}(\theta) < 1 \\ 1, & \text{otherwise.} \end{cases} \tag{6}$$

The stop-gradient operator $\mathrm{sg}[\cdot]$ ensures that $\mathcal{F}$ serves strictly as a confidence gate, effectively preventing the optimization objective from drifting. Mathematically, this transformation converts the optimization landscape from a disjoint cliff (characteristic of binary clipping) into a smooth and continuous manifold. Consequently, samples that marginally exceed the trust region contribute attenuated but non-zero gradients, ensuring that the

policy can still learn from "near-boundary" explorations without destabilizing the update dynamics.

4.2 Dual-Variable Adaptive Variance

The core innovation of MASPO lies in how the variance σ^2 is dynamically determined. We decouple σ into two components (Eq. (7) and (8)) to address the remaining two challenges: insensitive probability mass and asymmetric signal reliability.

$$\sigma_{\text{pos}} = \underbrace{\frac{\sigma_{\text{base}}}{\pi_{\theta_{old}}^{\alpha}}}_{\text{Mass-Adaptive}} \cdot \underbrace{\left(1 + \beta_{\text{high}} \hat{A}_{i,t}\right)}_{\text{Risk Controller}}, \quad (7)$$

$$\sigma_{\text{neg}} = \underbrace{\frac{\sigma_{\text{base}}}{\pi_{\theta_{old}}^{\alpha}}}_{\text{Mass-Adaptive}} \cdot \underbrace{\left(1 - \beta_{\text{low}} \hat{A}_{i,t}\right)^{-1}}_{\text{Risk Controller}}. \quad (8)$$

The Mass-Adaptive Limiter. The term $\frac{\sigma_{\text{base}}}{\pi_{\theta_{old}}^{\alpha}}$ serves as the **Mass-Adaptive Limiter**. It inversely scales the trust region width with the token's reference probability.

- **In the Long Tail ($\pi \to 0$):** The variance σ increases, widening the gate. This compensates for the negligible mass displacement of rare tokens, effectively expanding the exploration budget where it is most needed.
- **In the Head ($\pi \to 1$):** The variance shrinks, enforcing strict constraints to prevent policy collapse into a deterministic state due to excessive mass shifts. This aligns with the intuition that for high-confidence tokens, even small ratio deviations imply large absolute mass displacement, necessitating tighter variance to preserve the stability of the generation distribution.

The Asymmetric Risk Controller. The second component modulates the trust region based on the feedback signal, functioning as an **Asymmetric Risk Controller**. We derive this design from the monotonicity properties of group-relative advantages (see Appendix B), which link advantage magnitude to query difficulty.

- **Positive Signal Expansion (σ_{pos}):** For correct reasoning paths ($\hat{A} > 0$), we aim to maximize the utilization of verified signals. Based on the monotonicity property, difficult queries (where correct responses are rare) yield significantly larger positive advantages. These are high-value learning signals. Our expansion term $(1 + \beta_{\text{high}} \hat{A}_{i,t})$ sharpens this effect, allowing the model to take aggressively larger update steps for these high-confidence successes, accelerating the acquisition of complex reasoning skills.
- **Negative Signal Conservatism (σ_{neg}):** For incorrect paths ($\hat{A} < 0$), the handling is nuanced. According to GRPO monotonicity, "easy" queries (where most responses are correct) result in negative advantages with very large magnitudes (e.g., $A^- \ll -1$). In such easy tasks, a negative response is often a "near-miss" anomaly where the reasoning chain largely overlaps with correct paths but fails due to a minor error. Applying a massive penalty to these chains risks destroying valid intermediate reasoning patterns crucial for generalization.

  MASPO strategically addresses this via the compression term $(1-\beta_{\text{low}} \hat{A}_{i,t})^{-1}$. When $\hat{A}$ is large and negative, this term significantly shrinks σ_{neg}. This tightly constrains the update for failures in easy tasks, preventing "catastrophic unlearning" due to credit assignment ambiguity. Conversely, for hard tasks where failures are common (small negative advantage), the constraint remains relaxed, allowing normal optimization to proceed without excessive restriction.

5 Experiments

5.1 Experimental Setup

Implementation Details. We leverage the DeepSeek-R1-Distill-Qwen series (specifically 1.5B, 7B, and 14B) as our backbone models to assess MASPO. All models are fine-tuned on the DAPO-Math-17K dataset (Yu et al., 2025). For reward calculation, we adopt a dual-system approach: `math_verify` (Kydlíček, 2024) is employed for rule-based reward generation during training, whereas `prime_math` (Lightman et al., 2023a,b) serves as the evaluation metric.

Regarding the optimization dynamics, we configure a global batch size of 512 with a mini-batch size of 32. This setup facilitates 16 off-policy updates per importance sampling step, effectively amplifying the impact of off-policy updates. The clipping thresholds ε_{low} and $\varepsilon_{\text{high}}$ are both set to 0.2. Additional details are provided in the Appendix C.3.

Evaluation Protocols. To rigorously evaluate the generalization of the model's reasoning capabilities, we conduct evaluations across a diverse set of benchmarks, including AIME24 (MAA, 2024),

Table 1: Main results on mathematical reasoning benchmarks. We report Avg@32 (A@32) and Pass@32 (P@32). **Bold** indicates the best performance, and underlined represents the second-best.

	AIME24		AIME25		AMC23		MATH500		Minerva		Olympiad		Average	
Method	A@32	P@32	A@32	P@32	A@32	P@32	A@32	P@32	A@32	P@32	A@32	P@32	A@32	P@32
Backbone: DeepSeek-R1-Distill-Qwen-1.5B														
GRPO	33.2	71.8	27.7	49.9	79.5	94.8	77.6	90.8	26.1	48.8	46.3	64.7	48.4	70.1
Clip Higher	36.6	70.1	**30.1**	55.8	**82.8**	94.9	71.7	88.6	24.8	49.3	42.0	61.6	48.0	70.1
DAC	40.0	73.6	28.7	51.2	80.2	**95.0**	78.1	**92.5**	27.4	54.0	46.3	64.5	50.1	71.8
Entropy Adv.	29.6	67.8	23.9	44.2	77.1	94.1	78.5	89.7	25.2	48.6	46.4	64.3	46.8	68.1
BAPO	38.3	72.3	28.0	55.7	80.1	94.9	74.9	91.0	25.1	44.7	44.0	61.3	48.4	70.0
SAPO	39.3	73.7	28.0	49.9	82.7	94.8	**79.1**	91.4	29.6	52.7	**48.2**	**66.7**	51.2	71.5
MASPO	**41.0**	**74.8**	28.4	**58.0**	82.2	**95.0**	78.0	89.7	**30.7**	**54.1**	47.8	65.7	**51.4**	**72.9**
Backbone: DeepSeek-R1-Distill-Qwen-7B														
GRPO	48.2	**82.5**	37.4	60.5	88.1	96.6	84.8	92.4	37.4	57.2	57.2	73.9	58.9	77.2
Clip Higher	47.4	82.4	37.4	67.2	89.0	96.6	85.3	**95.4**	38.8	58.5	57.5	75.5	59.2	79.3
DAC	50.2	81.2	36.2	61.7	88.1	**99.6**	85.1	94.8	36.6	**58.7**	57.9	**76.0**	59.0	78.7
Entropy Adv.	49.1	81.5	34.4	58.1	87.7	96.6	84.8	92.5	36.5	54.0	56.4	73.7	58.2	76.1
BAPO	47.1	80.3	37.7	58.4	89.2	97.2	85.0	94.5	38.3	57.5	57.1	74.5	59.1	77.1
SAPO	47.5	79.2	35.3	58.0	88.7	95.0	85.5	92.9	**39.4**	58.1	56.6	74.6	58.8	76.3
MASPO	**53.2**	82.4	**42.9**	**73.2**	**91.4**	95.0	**86.0**	94.7	39.3	58.6	**58.0**	74.9	**61.8**	**79.8**

AIME25 (MAA, 2025), AMC23 (MAA, 2023), Minerva (Lewkowycz et al., 2022), OlympiadBench (He et al., 2024), and MATH500 (Hendrycks et al., 2021). We report performance using Avg@32 and Pass@32 metrics, which reflect the model's expected stability and its potential peak performance, respectively.

Baseline Methods. We benchmark MASPO against a comprehensive suite of strong methods, categorized by their optimization focus: (1) the original **GRPO** algorithm; (2) **SAPO**, which removes clipping boundaries entirely and uses a soft weighting function; (3) methods focusing on low-probability compensation, including **Entropy Regularization** and **DAC**; (4) approaches emphasizing positive sample rebalancing, such as **Advantage Reweighting**, **Entropy Advantage**, and **BAPO**; and (5) **Clip Higher**, which implicitly addresses both optimization directions. The comparative results are summarized in Table 1. Note that Entropy Regularization was only preliminarily tested on the 1.5B model, and Advantage Reweighting yielded suboptimal performance; consequently, detailed results for these two methods are omitted from the main table but are provided in Appendix D.1.

5.2 Main Results

Overall Performance Table 1 presents the comparative performance of MASPO against various baselines. MASPO delivers superior performance, achieving well-performing results on the aggregated average across both 1.5B and 7B scales. Specifically, on the 1.5B model, MASPO improves Avg@32 by 3.0% over GRPO and edges out the closest competitor (SAPO) by 0.2%. Crucially, this lead expands at the 7B scale. While other baselines plateau near GRPO levels – with the second-best Clip Higher leading GRPO by only 0.3% – MASPO establishes a clear margin, outperforming Clip Higher by 2.6% in Avg@32. Finally, Pass@K analysis (Figure 3(a,e)) confirms our method's robustness across difficulty levels: MASPO consistently dominates, leading on AIME24 with the 1.5B model and extending this superiority to the harder AIME25 benchmark at the 7B scale.

Training Dynamics. The training curves in Figure 3 reveal distinct optimization behaviors. GRPO exhibits a rapid decay in policy entropy, indicating an extreme exploitation mode that sacrifices exploration, consequently limiting its performance ceil-

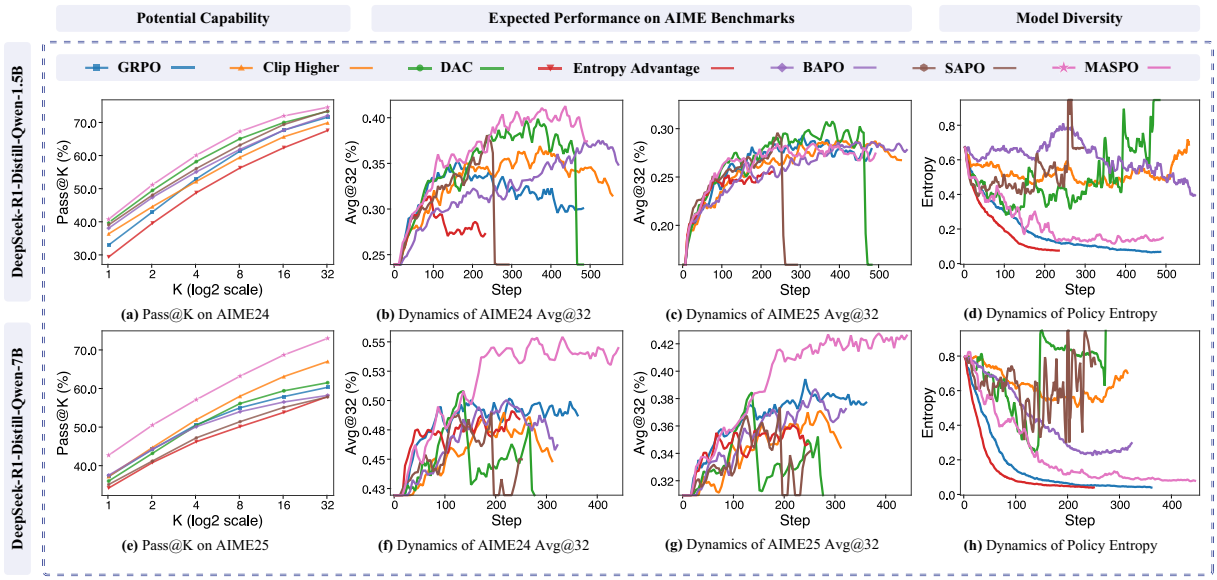


Figure 3: Evolution of Training Dynamics and Performance across Model Scales. Top row: 1.5B model; Bottom row: 7B model. MASPO demonstrates superior convergence, achieving higher performance ceilings.

ing. Entropy Advantage shows an even faster entropy drop, resulting in inferior final performance.

Among other baselines, Clip Higher, DAC, BAPO, and SAPO maintain higher entropy levels than GRPO, yet this does not strictly translate to better performance. Clip Higher and BAPO suffer from slow policy improvement, only surpassing GRPO on the 1.5B scale. SAPO experiences training collapse on both scales; it only exceeds GRPO's ceiling on the 1.5B model before diverging. We hypothesize this instability stems from its bilateral Gaussian gating design (see the unilateral ablation in Appendix D.2). DAC significantly outperforms GRPO on 1.5B but exhibits a "staircase" entropy spike on 7B, leading to performance degradation, though it still marginally beats GRPO before regression. In contrast, MASPO maintains a healthy entropy level higher than GRPO without instability or stagnation, achieving a performance ceiling significantly superior to all baselines.

5.3 Analysis of Hyperparameters

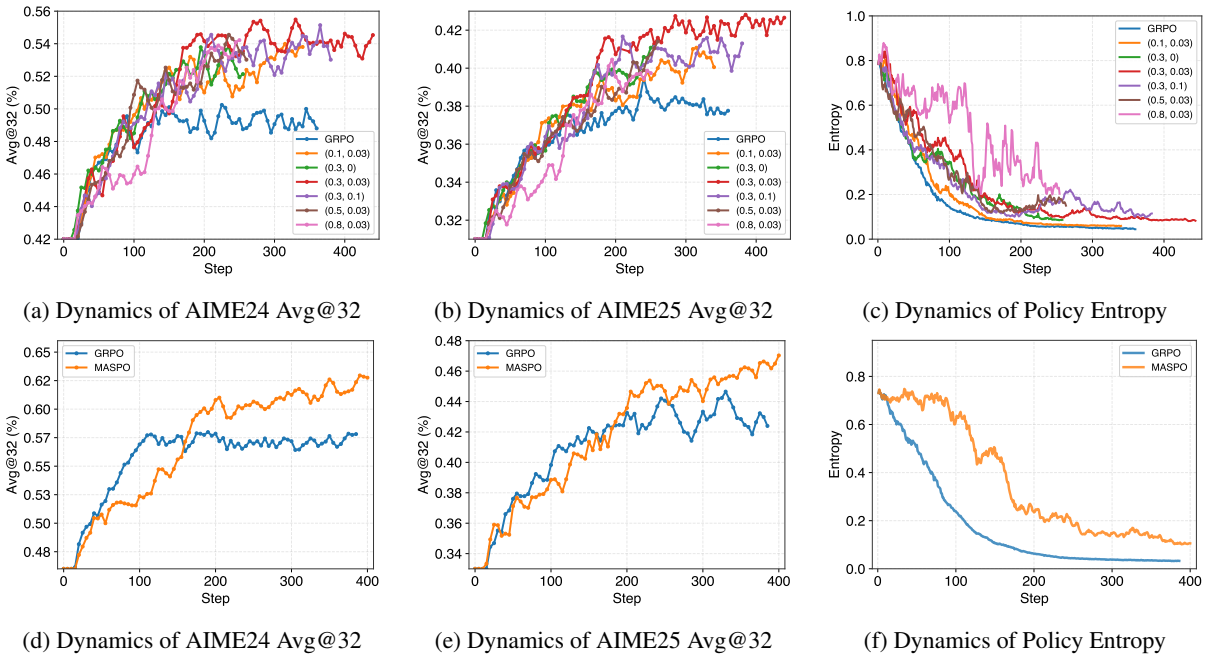
MASPO introduces four primary hyperparameters: σ_{base}, α, β_{low}, and β_{high}. To simplify the analysis, we fix $\sigma_{\text{base}} = 1$ and apply symmetric risk control ($\beta = \beta_{\text{low}} = \beta_{\text{high}}$). Using $\alpha = 0.3$, $\beta = 0.03$ as the baseline configuration, we vary parameters to test the following set: $(0.1, 0.03)$, $(0.3, 0)$, $(0.3, 0.03)$, $(0.3, 0.1)$, $(0.5, 0.03)$, $(0.8, 0.03)$. Experiments were conducted on the 7B model, with results shown in the top row of Figure 4 and Table 5. Our analysis yields the following conclusions (comprehensive data in Appendix D.3):

Robustness. As evidenced by the curves in Figure 4 and Appendix D.3, MASPO demonstrates exceptional robustness. Configurations with $\alpha \in \{0.3, 0.5\}$ yield the optimal performance ceilings. While extreme values ($\alpha = 0.1$ or 0.8) lead to a slight drop, they still consistently outperform GRPO. Furthermore, varying β across the tested range maintains superiority over the baseline.

Impact of Mass-Adaptive Scaling (α). Setting $\alpha = 0.1$ imposes excessive constraints, resulting in low entropy levels similar to GRPO and premature loss of exploration capability. Conversely, setting $\alpha = 0.8$ overly relaxes constraints, leading to entropy instability and degraded performance. Based on supplementary experiments in Appendix D.3, we recommend $\alpha \in [0.3, 0.5]$ as the optimal range for balancing stability and exploration.

Impact of Asymmetric Risk Control (β). Using the baseline $\alpha = 0.3$, setting $\beta = 0$ (disabling risk control) causes entropy to drop too rapidly in the early stages, curtailing exploration and slightly lowering the performance ceiling. Increasing β to 0.1 yields a similar effect of premature entropy decay. We recommend $\beta = 0.03$ as the optimal setting when $\alpha = 0.3$.

While the efficacy of risk control on negative samples is theoretically justified by the misalignment between advantage and signal-to-noise ratio (see derivation in Appendix B), the benefit of sharpening positive advantages requires empirical validation. Our decoupling experiment (Appendix D.4) confirms that applying risk control to positive sam-



(a) Dynamics of AIME24 Avg@32 (b) Dynamics of AIME25 Avg@32 (c) Dynamics of Policy Entropy

(d) Dynamics of AIME24 Avg@32 (e) Dynamics of AIME25 Avg@32 (f) Dynamics of Policy Entropy

Figure 4: Hyperparameter sensitivity and scalability analysis. **Top:** MASPO demonstrates robustness across a range of α and β values, outperforming the GRPO baseline. **Bottom:** Training dynamics on the larger 14B model, demonstrating MASPO's scalability by maintaining higher entropy and superior performance compared to GRPO.

ples is crucial for sustaining long-term exploration and preventing performance regression.

5.4 Scaling Analysis

To further investigate the scalability of our approach, we applied MASPO to the larger DeepSeek-R1-Distill-Qwen-14B model using $\alpha = 0.5$ and $\beta = 0.03$. The bottom row of Figure 4 illustrates the training dynamics. MASPO maintains higher entropy and achieves a significant performance margin over the GRPO baseline. The relative improvements of MASPO across different model scales are summarized in Table 2 (see Appendix D.5 for more details). It is worth noting that the 1.5B and 7B versions are distilled from Qwen-2.5-Math, providing a stronger initial mathematical capability compared to the 14B version, which is based on Qwen-2.5-General. This distinction accounts for the baseline performance variance observed in our scaling experiments.

Table 2: Scalability analysis of average Avg@32 (A) and Pass@32 (P) across 1.5B, 7B, and 14B models.

Scale	GRPO		MASPO		Gain ↑	
	A	P	A	P	A	P
1.5B	48.4	70.1	**51.4**	**72.9**	+3.0	+2.8
7B	58.9	77.2	**61.8**	**79.8**	+2.9	+2.6
14B	53.6	67.4	**56.4**	**71.1**	+2.8	+3.7

6 Conclusion

In this work, we identify and address structural disconnects in current RLVR paradigms, specifically the inefficiencies from rigid, uniform, and symmetric trust region constraints. We argue prevailing "hard clipping" mechanisms fail to accommodate complex dynamics of LLM optimization, particularly regarding the heavy-tailed nature of token distributions and the disparate reliability of reward signals. To bridge these gaps and establish a unified theoretical framework, we propose **M**ass-**A**daptive **S**oft **P**olicy **O**ptimization (MASPO).

By unifying a differentiable Soft Gaussian Gating mechanism with a Mass-Adaptive Limiter and an Asymmetric Risk Controller, MASPO creates a continuous and responsive optimization landscape. This framework ensures that exploration budgets are dynamically allocated based on token probability mass and that update magnitudes are modulated by the signal-to-noise ratio of the verification feedback. Extensive evaluations across the DeepSeek-R1-Distill-Qwen series (1.5B, 7B, and 14B) confirm that MASPO significantly enhances sample efficiency and reasoning accuracy. Our findings suggest that moving beyond rigid box constraints toward adaptive, probability-aware, and risk-sensitive optimization is a critical step for the next generation of reasoning models.

Limitations

We must acknowledge two primary limitations in our current study. (1) **Dependence on Verifiable Rewards:** Our Asymmetric Risk Controller relies on the premise that positive rewards represent verified truths while negative rewards contain ambiguity (credit assignment noise). Consequently, our experiments are concentrated on mathematical reasoning tasks (e.g., AIME, MATH500) where ground truth is deterministic. The applicability of MASPO to domains with subjective or partial rewards (such as creative writing, general chat, or multimodal retrieval) (Li et al., 2026; Chen et al., 2025b, 2026) requires further adaptation of the risk control mechanism. (2) **Computational Scope:** While we demonstrate scalability from 1.5B to 14B parameters, we have not yet verified MASPO on extremely large-scale foundation models (e.g., 70B+ or MoE architectures) due to computational resource constraints. Although the theoretical derivation of mass-adaptive scaling suggests robustness, empirical validation on such scales remains an important avenue for our future research work.

Ethical Considerations

We have carefully considered potential societal and the ethical implications of our research:

- **Research Integrity:** Throughout this study, we have adhered to established ethical guidelines. We ensure that our methodology is reported transparently, with full disclosure of theoretical derivations and implementation details.
- **Data Usage and Privacy:** The datasets utilized in our experiments (e.g., DAPO-Math-17k, AIME) are derived from publicly available, peer-reviewed scientific sources. No private, sensitive, or personally identifiable information (PII) was processed or generated during the course of this research. This commitment to responsible data practices is consistent with broader research efforts (Shao et al., 2025a,b).
- **Reproducibility:** To foster scientific progress and transparency within the community, we provide comprehensive documentation of our hyperparameter configurations, including the specific tuning ranges for α and β.
- **Open Access:** In the interest of community collaboration, we have made our code available anonymously for review and commit to fully open-sourcing the complete implementation upon the final acceptance of this paper.

References

Sikai Bai, Haoxi Li, Jie Zhang, Yongjiang Liu, and Song Guo. 2026. Ttvs: Boosting self-exploring reinforcement learning via test-time variational synthesis. *arXiv preprint arXiv:2604.08468*.

Aili Chen, Aonian Li, Bangwei Gong, Binyang Jiang, Bo Fei, Bo Yang, Boji Shan, Changqing Yu, Chao Wang, Cheng Zhu, and 1 others. 2025a. Minimax-m1: Scaling test-time compute efficiently with lightning attention. *arXiv preprint arXiv:2506.13585*.

Zhiwei Chen, Yupeng Hu, Zhiheng Fu, Zixu Li, Jiale Huang, Qinlei Huang, and Yinwei Wei. 2026. Intent: Invariance and discrimination-aware noise mitigation for robust composed image retrieval. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 20463–20471.

Zhiwei Chen, Yupeng Hu, Zixu Li, Zhiheng Fu, Xuemeng Song, and Liqiang Nie. 2025b. Offset: Segmentation-based focus shift revision for composed image retrieval. In *ACM International Conference on Multimedia (ACM MM)*, pages 6113–6122.

Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Wayne Xin Zhao, Zhenliang Zhang, and Furu Wei. 2025. Reasoning with exploration: An entropy perspective. *arXiv preprint arXiv:2506.14758*.

Haonan Dong, Kehan Jiang, Haoran Ye, Wenhao Zhu, Zhaolu Kang, and Guojie Song. 2026. Neureasoner: Towards explainable, controllable, and unified reasoning via mixture-of-neurons. *Preprint*, arXiv:2604.02972.

Haonan Dong, Wenhao Zhu, Guojie Song, and Liang Wang. 2025. AuroRA: Breaking low-rank bottleneck of loRA with nonlinear mapping. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Benjamin Eysenbach and Sergey Levine. 2021. Maximum entropy rl (provably) solves some robust rl problems. *arXiv preprint arXiv:2103.06257*.

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. *arXiv preprint arXiv:2511.20347*.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, and 1 others. 2024. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*.

Jujie He, Jiacai Liu, Chris Yuhao Liu, Rui Yan, Chaojie Wang, Peng Cheng, Xiaoyu Zhang, Fuxiang Zhang,

Jiacheng Xu, Wei Shen, and 1 others. 2025. Skywork open reasoner 1 technical report. *arXiv preprint arXiv:2505.22312*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Guanhua Huang, Tingqiang Xu, Mingze Wang, Qi Yi, Xue Gong, Siheng Li, Ruibin Xiong, Kejiao Li, Yuhao Jiang, and Bo Zhou. 2025. Low-probability tokens sustain exploration in reinforcement learning with verifiable reward. *arXiv preprint arXiv:2510.03222*.

Edwin T Jaynes. 1957a. Information theory and statistical mechanics. *Physical review*, 106(4):620.

Edwin T Jaynes. 1957b. Information theory and statistical mechanics. ii. *Physical review*, 108(2):171.

Kehan Jiang, Haonan Dong, Zhaolu Kang, Zhengzhou Zhu, and Guojie Song. 2026. Foe: Forest of errors makes the first solution the best in large reasoning models. *Preprint*, arXiv:2604.02967.

Hynek Kydlíček. 2024. Math-Verify: Math Verification Library.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, and 1 others. 2022. Solving quantitative reasoning problems with language models. *Advances in neural information processing systems*, 35:3843–3857.

Zixu Li, Yupeng Hu, Zhiwei Chen, Shiqi Zhang, Qinlei Huang, Zhiheng Fu, and Yinwei Wei. 2026. Habit: Chrono-synergia robust progressive learning framework for composed image retrieval. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 6762–6770.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023a. Let's verify step by step. *arXiv preprint arXiv:2305.20050*.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023b. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

Peiyang Liu. 2024. Unsupervised corrupt data detection for text training. *Expert Systems with Applications*, 248:123335.

Peiyang Liu, Ziqiang Cui, Di Liang, and Wei Ye. 2025. Who stole your data? a method for detecting unauthorized rag theft. *arXiv preprint arXiv:2510.07728*.

Peiyang Liu, Jinyu Yang, Lin Wang, Sen Wang, Yunlai Hao, and Huihui Bai. 2023. Retrieval-based unsupervised noisy label detection on text data. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 4099–4104.

MAA. 2023. American mathematics competitions - amc. Accessed: 2023.

MAA. 2024. American invitational mathematics examination - aime. Accessed: 2024.

MAA. 2025. American invitational mathematics examination - aime. Accessed: 2025.

Brendan O'Donoghue, Remi Munos, Koray Kavukcuoglu, and Volodymyr Mnih. 2016. Combining policy gradient and q-learning. *arXiv preprint arXiv:1611.01626*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Doina Precup, Richard S Sutton, and Satinder P Singh. 2000. Eligibility traces for off-policy policy evaluation. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 759–766.

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

Zekai Shao, Yi Shan, Yixuan He, Yuxuan Yao, Junhong Wang, Xiaolong Zhang, Yu Zhang, and Siming Chen. 2025a. Do language model agents align with humans in rating visualizations? an empirical study. *IEEE Computer Graphics and Applications*.

Zekai Shao, Leixian Shen, Haotian Li, Yi Shan, Huamin Qu, Yun Wang, and Siming Chen. 2025b. Narrative player: Reviving data narratives with visuals. *IEEE Transactions on Visualization and Computer Graphics*, 31(10):6781–6795.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

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

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

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

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. 2022. Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*.

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

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

Ronald J Williams and Jing Peng. 1991. Function optimization using connectionist reinforcement learning algorithms. *Connection Science*, 3(3):241–268.

Zhiheng Xi, Xin Guo, Yang Nan, Enyu Zhou, Junrui Shen, Wenxiang Chen, Jiaqi Liu, Jixuan Huang, Zhihao Zhang, Honglin Guo, and 1 others. 2025. Bapo: Stabilizing off-policy reinforcement learning for llms via balanced policy optimization with adaptive clipping. *arXiv preprint arXiv:2510.18927*.

Wei Xiong, Jiarui Yao, Yuhui Xu, Bo Pang, Lei Wang, Doyen Sahoo, Junnan Li, Nan Jiang, Tong Zhang, Caiming Xiong, and 1 others. 2025. A minimalist approach to llm reasoning: from rejection sampling to reinforce. *arXiv preprint arXiv:2504.11343*.

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

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

Deheng Ye, Zhao Liu, Mingfei Sun, Bei Shi, Peilin Zhao, Hao Wu, Hongsheng Yu, Shaojie Yang, Xipeng Wu, Qingwei Guo, and 1 others. 2020. Mastering complex control in moba games with deep reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, pages 6672–6679.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, and 1 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.

Naixin Zhai, Pengyang Shao, Binbin Zheng, Yonghui Yang, Fei Shen, Long Bai, and Xun Yang. 2026. Maximizing local entropy where it matters: Prefix-aware localized llm unlearning. *arXiv preprint arXiv:2601.03190*.

A Theoretical Motivation and Derivation

In this appendix, we provide the rigorous theoretical justification for the **M**ass-**A**daptive **S**oft **P**olicy **O**ptimization (MASPO) objective. We deliberately deviate from the standard KL-divergence-based derivation used in TRPO and PPO, arguing that a framework combining *Probability Mass Displacement* with *Probabilistic Relaxation* offers a more robust and mathematically consistent foundation for fine-tuning Large Language Models.

A.1 Step 1: Metric Selection

Standard policy optimization methods constrain the update using the Kullback-Leibler (KL) divergence, $D_{KL}(\pi_{\theta_{\text{old}}}||\pi_\theta)$. While theoretically sound for distribution matching, KL divergence measures the relative information gain, which can be insensitive to the absolute probability mass of tokens.

In the context of LLM fine-tuning, the stability of the generation process is better preserved by bounding the absolute shift in probability assignment. Ideally, this corresponds to the Total Variation (TV) distance:

$$D_{TV}(\pi_\theta, \pi_{\theta_{\text{old}}}) = \frac{1}{2}\sum_x |\pi_\theta(x) - \pi_{\theta_{\text{old}}}(x)|. \quad (9)$$

However, optimizing the L_1 norm directly is computationally challenging due to its non-differentiability at zero. Consequently, we adopt the **Squared Mass Displacement** (an L_2 proxy) as our constraint metric. For a specific token transition, we require the policy shift to satisfy:

$$\frac{1}{2}(\pi_\theta - \pi_{\theta_{\text{old}}})^2 \leq \delta^2, \quad (10)$$

where δ is a localized trust region budget. By substituting the probability ratio $\rho(\theta) = \pi_\theta/\pi_{\theta_{\text{old}}}$, we can rewrite this constraint in terms of the ratio:

$$\begin{aligned} &\frac{1}{2}\pi^2_{\theta_{\text{old}}}(\rho(\theta) - 1)^2 \leq \delta^2 \\ &\implies (\rho(\theta) - 1)^2 \leq \frac{2\delta^2}{\pi^2_{\theta_{\text{old}}}}. \end{aligned} \quad (11)$$

Theoretical Implication: This derivation reveals a fundamental and critical insight: to maintain a constant safety margin in terms of probability mass (δ),

the permissible deviation of the ratio $(\rho - 1)$ must be *inversely proportional* to the reference probability $\pi_{\theta_{\text{old}}}$. This finding theoretically justifies the base scaling term in MASPO:

$$\sigma_{\text{base}} \propto \frac{1}{\pi_{\theta_{\text{old}}}^{\alpha}}. \tag{12}$$

While the strict derivation suggests $\alpha = 1$, we introduce α as a hyperparameter to empirically control the sensitivity to the long-tail distribution, acknowledging that the L_2 proxy may over-encourage rare tokens in high-dimensional vocabularies.

A.2 Step 2: MaxEnt-based Soft Relaxation

A core technical challenge in GRPO is the use of "hard clipping", which effectively imposes a rigid box constraint on the ratio. This creates a disjoint optimization landscape with zero gradients vanishing outside the clip range. In MASPO, we reformulate the constrained optimization problem as a **Confidence-Weighted Estimation** problem.

Let the "cost" of a policy update for a specific sample be defined by mass displacement violation:

$$\mathcal{C}(\rho) = \frac{(\rho - 1)^2}{2\sigma^2}, \tag{13}$$

where σ is the adaptive variance previously derived in Step 1. Instead of enforcing a rigid hard constraint $\mathcal{C}(\rho) \leq C$, we interpret this cost as a direct measure of **Gradient Untrustworthiness**. We seek a weighting function $w(\rho) \in [0, 1]$ that modulates the gradient contribution based on this cost.

According to the **Principle of Maximum Entropy** (Jaynes, 1957a,b; Zhai et al., 2026), if we possess information about the expected cost (first moment) but are otherwise ignorant about the distribution, the entropy-maximizing distribution is the exponential family. Specifically, we formulate the problem as finding a weight distribution $w(\rho)$ that maximizes entropy $H(w)$ subject to an expected cost constraint:

$$\begin{aligned} \max_{w} \quad & H(w) = -\int w(\rho) \log w(\rho) d\rho \\ \text{s.t.} \quad & \mathbb{E}_w[\mathcal{C}(\rho)] = \mu. \end{aligned} \tag{14}$$

The solution to this Lagrangian optimization is the Gaussian (or Radial Basis) kernel:

$$w^*(\rho) \propto \exp\left(-\beta \cdot \mathcal{C}(\rho)\right) = \exp\left(-\frac{(\rho - 1)^2}{2\sigma^2}\right). \tag{15}$$

Here, the Lagrange multiplier β is absorbed into σ.

From Distribution to Gradient Weighting: We apply this optimal distribution $w^*(\rho)$ as a soft gating mechanism for the policy gradient. Unlike the standard objective which assumes uniform reliability across all samples, MASPO optimizes a *reliability-weighted* lower bound:

$$\nabla \mathcal{J}_{\text{MASPO}} \approx \mathbb{E}_{o \sim \pi}\Bigg[\underbrace{\exp\left(-\frac{(\rho - 1)^2}{2\sigma^2}\right)}_{\text{Soft Trust Region}} \cdot \rho(\theta) \hat{A} \cdot \nabla \log \pi_\theta \Bigg]. \tag{16}$$

This formulation smoothly down-weights samples that exhibit excessive mass displacement (high variance/low confidence) rather than abruptly discarding them, thereby providing a more differentiable and continuous optimization manifold.

A.3 Step 3: Asymmetric Risk Modeling

The final component of the derivation addresses the dynamic nature of the budget σ. While Step 1 established the dependency on probability mass $\pi_{\theta_{\text{old}}}$, we must also account for the nature of the reward signal in reasoning tasks.

Unlike the standard assumption in TRPO/PPO where advantage estimates are treated symmetrically, we proceed from the empirical observation that the Signal-to-Noise Ratio (SNR) in mathematical reasoning and code generation is highly asymmetric. A positive reward verifies a correct reasoning path, while a negative reward is ambiguous (credit assignment problem). We model this asymmetry by modulating the "trust budget" σ based on the sign of the advantage $\hat{A}$.

Case 1: Positive Advantage (High SNR). When $\hat{A} > 0$, the sample indicates a verifiable success. Empirically, we want to exploit these high-quality signals aggressively. We model this as a linear expansion of the trust region:

$$\sigma_{\text{pos}}(\hat{A}) \propto \frac{1}{\pi_{\theta_{\text{old}}}^{\alpha}} \cdot (1 + \beta_{\text{high}} \hat{A}). \tag{17}$$

This formulation allows the policy to take larger steps when the signal is strong and positive, akin to an "Optimism in the Face of Uncertainty" strategy.

Case 2: Negative Advantage (Low SNR). When $\hat{A} < 0$, the signal is often noisy. An overly aggressive penalty based on a potentially correct intermediate step (labeled negative due to a later

error) can lead to unrecoverable policy collapse. Therefore, we treat the negative regime much more conservatively. We use an inverse scaling to tighten the boundary as the negative advantage grows:

$$\sigma_{\text{neg}}(\hat{A}) \propto \frac{1}{\pi^{\alpha}_{\theta_{\text{old}}}} \cdot \frac{1}{1-\beta_{\text{low}}\hat{A}}. \tag{18}$$

This acts as a safeguard, restricting the magnitude of policy shifts driven by noisy negative feedback.

Theoretical Alignment: While derived from the specific characteristics of CoT data, this asymmetric variance design aligns with **Risk-Sensitive Control** theory. It effectively assigns a higher risk cost to updates driven by negative samples (tightening the constraint) while lowering the cost for positive samples (relaxing the constraint), thereby optimizing the exploration-exploitation trade-off in sparse-reward environments.

In the implementation, we use clipping for numerical stability:

$$\sigma_{\text{pos}} = \min\left(\frac{\sigma_{\text{base}}}{\pi^{\alpha}_{\theta_{old}}}, 10\right) \cdot \text{clip}\left(1+\beta_{\text{high}}\hat{A}_{i,t}, 0.1, 10\right), \tag{19}$$

$$\sigma_{\text{neg}} = \min\left(\frac{\sigma_{\text{base}}}{\pi^{\alpha}_{\theta_{old}}}, 10\right) \cdot \text{clip}\left(\frac{1}{1-\beta_{\text{low}}\hat{A}_{i,t}}, 0.1, 10\right). \tag{20}$$

B Monotonicity Analysis of Group Relative Advantages

In this section, we analyze the relationship between the advantage values and the difficulty of the task, defined by the number of correct samples within a group. This analysis theoretically supports the asymmetric variance design in MASPO.

Consider a group of size n, with x correct samples (positive rewards $r=1$) and $n-x$ incorrect samples (negative rewards $r=-1$).

First, we calculate the mean reward μ:

$$\mu = \frac{1}{n}\left[x(1)+(n-x)(-1)\right] = \frac{2x-n}{n}. \tag{21}$$

Next, we calculate the variance σ^2. Note that since $r \in \{-1,1\}$, r^2 is always 1. Thus, the second moment $\mathbb{E}[r^2]=1$. The variance is:

$$\begin{aligned}\sigma^2 &= \mathbb{E}[r^2] - (\mathbb{E}[r])^2 = 1 - \left(\frac{2x-n}{n}\right)^2 \\ &= 1-(2p-1)^2 = 1-(4p^2-4p+1) \\ &= 4p(1-p).\end{aligned} \tag{22}$$

where $p = x/n$ is the accuracy rate. The standard deviation is $\sigma = 2\sqrt{p(1-p)} = \frac{2\sqrt{x(n-x)}}{n}$.

The advantages for positive (A^+) and negative (A^-) samples are standardized as follows.

For positive samples ($r=1$):

$$\begin{aligned}A^+(x) &= \frac{1-\mu}{\sigma} = \frac{1-(2p-1)}{2\sqrt{p(1-p)}} \\ &= \frac{2(1-p)}{2\sqrt{p(1-p)}} = \sqrt{\frac{1-p}{p}} \\ &= \sqrt{\frac{n-x}{x}}.\end{aligned} \tag{23}$$

For negative samples ($r=-1$):

$$\begin{aligned}A^-(x) &= \frac{-1-\mu}{\sigma} = \frac{-1-(2p-1)}{2\sqrt{p(1-p)}} \\ &= \frac{-2p}{2\sqrt{p(1-p)}} = -\sqrt{\frac{p}{1-p}} \\ &= -\sqrt{\frac{x}{n-x}}.\end{aligned} \tag{24}$$

Despite the shift in reward scale from $\{0,1\}$ to $\{-1,1\}$, the final expressions for A^+ and A^- remain invariant. We now analyze their monotonicity with respect to x (for $x \in (0,n)$):

Positive Advantage (A^+).

$$\begin{aligned}\frac{dA^+}{dx} &= \frac{d}{dx}\left(\frac{n}{x}-1\right)^{1/2} \\ &= -\frac{n}{2x^2}\sqrt{\frac{x}{n-x}} < 0.\end{aligned} \tag{25}$$

$A^+(x)$ is strictly decreasing with respect to x. This implies that for **hard tasks** (small x), the positive advantage is large. MASPO's expansion term $(1+\beta_{\text{high}}A^+)$ amplifies this, encouraging larger updates for these rare and valuable successes.

Negative Advantage (A^-).

$$\begin{aligned}\frac{dA^-}{dx} &= -\frac{d}{dx}\left(\frac{x}{n-x}\right)^{1/2} \\ &= -\frac{n}{2(n-x)^2}\sqrt{\frac{n-x}{x}} < 0.\end{aligned} \tag{26}$$

$A^-(x)$ is strictly decreasing (becoming more negative) as x increases. This implies that for **easy tasks** (large x), the negative advantage has a large absolute magnitude (e.g., $x \to n \implies A^- \to -\infty$). These samples are likely "near-misses" in an otherwise easy context. MASPO's compression term

$(1 - \beta_{\text{low}} A^-)^{-1}$ becomes very small when A^- is large and negative, effectively suppressing the gradient to prevent excessive penalization of these noisy and low-credit negative samples.

C Implementation Details

C.1 Experimental Ecosystem

Training Infrastructure. All experiments were orchestrated on a distributed high-performance computing cluster. The hardware topology consists of 30 compute nodes, each populated with 8 × NVIDIA A100 (80GB) GPUs. Inter-node communication is facilitated by a high-bandwidth InfiniBand fabric, while intra-node data transfer utilizes NVLink to minimize gradient synchronization latency. The cumulative compute budget for this study exceeded 20,000 GPU-hours.

Software Pipeline. We implemented MASPO within the VeRL framework (Sheng et al., 2025), optimized for post-training large-scale reasoning.

- **Distributed Strategy:** We employed Fully Sharded Data Parallel (FSDP) to shard parameters, gradients, and optimizer states across ranks. This allows full-parameter fine-tuning of the 14B model without necessitating parameter-efficient approximations like LoRA.

- **Inference Optimization:** To handle extensive rollout generation required by group-based RL, we integrated vLLM with PagedAttention. This setup efficiently manages the Key-Value (KV) cache, enabling high-throughput generation of long chain-of-thought sequences (up to 8192 tokens) with minimal memory fragmentation (Dong et al., 2025).

C.2 Task Definition and Evaluation

Training Data and Tokenization. Our training utilizes the DAPO-Math-17k dataset, which consists entirely of publicly available mathematical problems with verifiable ground-truth answers, ensuring data integrity throughout the pipeline (Liu et al., 2025). We strictly adhere to the Qwen2.5-Math tokenizer to maintain alignment between supervised fine-tuning (SFT) and the RLVR stage. We enforce a maximum sequence length of 8192 tokens, a critical setting that permits the model to explore complex, multi-step reasoning paths without artificial truncation.

Reward Mechanism. We adopt a sparse, outcome-based reward function. For each query q, the generated response o is parsed to extract the final answer. We utilize the `math_verify` library for robust answer extraction and equivalence checking.

$$r(q, o) = \begin{cases} 1, & \text{if answer matches ground truth} \\ -1, & \text{otherwise.} \end{cases} \tag{27}$$

This binary signal $\{-1, 1\}$ presents a challenging exploration landscape (Liu et al., 2023; Liu, 2024), necessitating the mass-adaptive and risk-sensitive mechanisms embedded within MASPO.

Benchmark Suite. To provide a holistic assessment of mathematical reasoning, we evaluate on a diverse set of benchmarks ranging from high school competitions to collegiate-level problems. Detailed characteristics are provided in Table 3.

C.3 Configuration and Reproducibility

Optimization Protocol. We utilize the AdamW optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.95$. To isolate the algorithmic contribution of MASPO, we employ a **constant learning rate** schedule across all experiments, avoiding the confounding effects of warm-up or decay schedules. The global batch size is set to 512 for rollouts and 32 for updates. We set the KL penalty coefficient $\beta_{\text{KL}} = 0$, relying implicitly on the trust region constraints.

Baseline Configurations. For all baseline algorithms, we adopt the optimal hyperparameter settings reported in their respective original papers or derived from our best-effort grid search (as detailed in Section D.1). Table 4 lists the specific configurations used for the main results in Table 1.

D Supplementary Experimental Analysis

D.1 Additional Baseline Experiments

We provide further context for baseline comparisons. For Clip Higher, the default $\varepsilon_{\text{high}} = 0.28$ performed poorly on both 1.5B and 7B scales (Figure 5). Following prior exploration strategies (He et al., 2025), we adopted $\varepsilon_{\text{high}} = 0.265$, which yielded the competitive results reported in the main Table 1. For Advantage Reweighting, the default $\alpha_{\text{A}} = 0.3$ underperformed; adjusting to $\alpha_{\text{A}} = 0.1$ improved results, though it still lagged behind GRPO. Experiments with Entropy Regularization on the 1.5B scale showed negligible difference compared to standard GRPO.

Table 3: Summary of the evaluation benchmarks used to assess model performance.

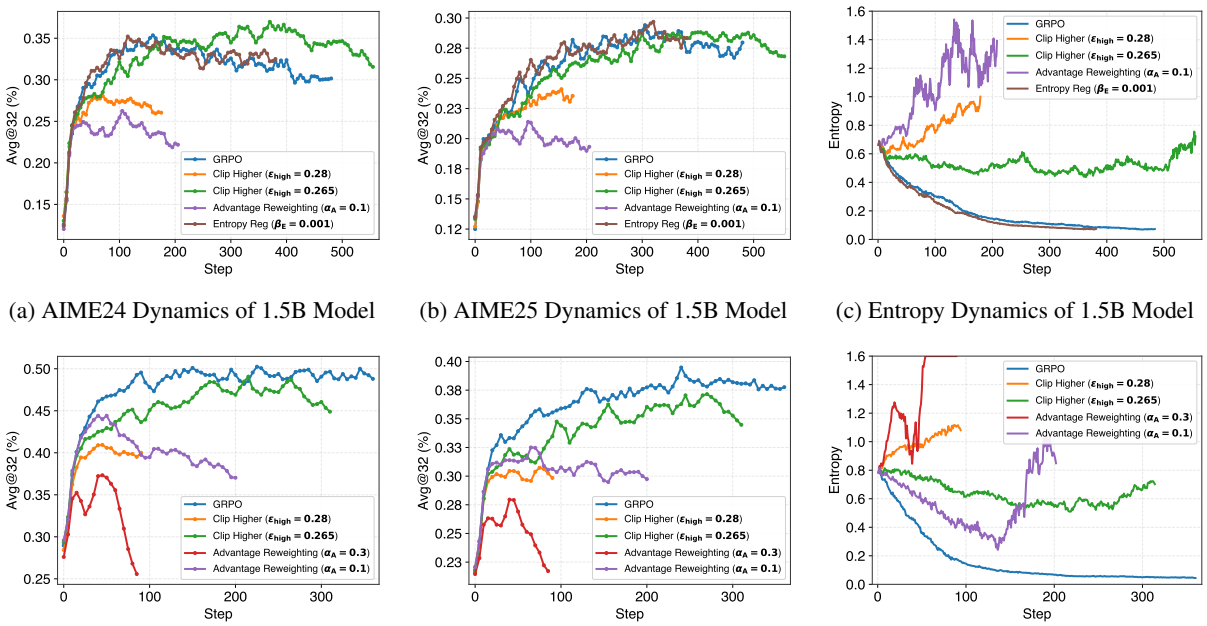
Benchmark	Evaluation Focus & Characteristics
AIME 2024	Focuses on arithmetic precision and robustness against contamination. Requires integer outputs (000–999).
AIME 2025	A strictly held-out set representing the frontier of current reasoning capabilities. Tests generalization to unseen, complex logical chains.
AMC 2023	Covers foundational algebra, geometry, and counting. Serves as a baseline for competitive mathematical proficiency.
MATH-500	A representative subset of the MATH dataset, spanning 7 domains including Calculus. Heavily tests LaTeX parsing and symbolic manipulation.
Minerva	Derived from scientific literature, testing higher-order reasoning and domain-specific notation understanding.
OlympiadBench	An aggregate of international competitions (IMO, CMO), representing the upper bound of cross-lingual mathematical logic.

Table 4: Detailed hyperparameter configurations for MASPO and all baselines. The settings listed here correspond to the results reported in the main experimental table.

Category	Algorithm	Learning Rate	Key Hyperparameters
Common Settings: Global Batch Size = 512, Mini-batch Size = 32, Max Length = 8192, $\beta_{KL} = 0$			
DeepSeek-R1-Distill-Qwen-1.5B	GRPO	1.0×10^{-6}	$\varepsilon = 0.2$
	Entropy Reg.		$\beta_{\text{E}} = 0.01$
	Clip Higher		$\varepsilon_{\text{low}} = 0.2, \varepsilon_{\text{high}} = 0.265$
	Adv. Reweighting		$\alpha_{\text{A}} = 0.1$
	Entropy Adv.		$\alpha_{\text{E}} = 0.4, \kappa = 2.0$
	DAC		$\varepsilon_{\text{low}} = 0.2, \varepsilon_{\text{high}} = 0.2$
	BAPO		$\rho_0 = 0.4, a^+ = 1.2, b^+ = 3.0, a^- = 0.6,$ $b^- = 0.9, \delta_1 = 0.05, \delta_2 = 0.02$
	SAPO		$\tau_{\text{neg}} = 1.05, \tau_{\text{pos}} = 1.0$
	MASPO		$\sigma_{\text{base}} = 1, \alpha = 0.5, \beta = 0.03$
DeepSeek-R1-Distill-Qwen-7B	GRPO	4.63×10^{-7}	$\varepsilon = 0.2$
	Clip Higher		$\varepsilon_{\text{low}} = 0.2, \varepsilon_{\text{high}} = 0.265$
	Adv. Reweighting		$\alpha_{\text{A}} = 0.1$
	Entropy Adv.		$\alpha_{\text{E}} = 0.4, \kappa = 2.0$
	DAC		$\varepsilon_{\text{low}} = 0.2, \varepsilon_{\text{high}} = 0.2$
	BAPO		$\rho_0 = 0.4, a^+ = 1.2, b^+ = 3.0, a^- = 0.6,$ $b^- = 0.9, \delta_1 = 0.05, \delta_2 = 0.02$
	SAPO		$\tau_{\text{neg}} = 1.05, \tau_{\text{pos}} = 1.0$
	MASPO		$\sigma_{\text{base}} = 1, \alpha = 0.3, \beta = 0.03$
DeepSeek-R1-Distill-Qwen-14B	GRPO	3.27×10^{-7}	$\varepsilon = 0.2$
	MASPO		$\sigma_{\text{base}} = 1, \alpha = 0.5, \beta_{\text{low}} = 0.03$

D.2 Ablation on SAPO Gating Mechanism

To investigate the instability of SAPO, we modified it to adopt a unilateral gating form identical to MASPO—applying the soft Gaussian gating only when the update is potentially destabilizing, i.e., when $(\hat{A}_{i,t} > 0 \wedge \rho_{i,t}(\theta) > 1)$ or $(\hat{A}_{i,t} < 0 \wedge \rho_{i,t}(\theta) < 1)$. Unlike SAPO's bilateral design, which indiscriminately attenuates all token updates and suppresses even conservative gradients, the unilateral form preserves beneficial updates. Experiments on DeepSeek-R1-Distill-Qwen-1.5B show that this unilateral version avoids the collapse observed in the original bilateral SAPO. This variant effectively serves as an ablation of



(a) AIME24 Dynamics of 1.5B Model (b) AIME25 Dynamics of 1.5B Model (c) Entropy Dynamics of 1.5B Model

(d) AIME24 Dynamics of 7B Model (e) AIME25 Dynamics of 7B Model (f) Entropy Dynamics of 7B Model

Figure 5: Supplementary training dynamics for additional baselines (Clip Higher, Advantage Reweighting, etc.).

MASPO without the Mass-Adaptive Variance and Risk Controller, further validating the rationale behind MASPO's unilateral design.

D.3 Complete Hyperparameter Results

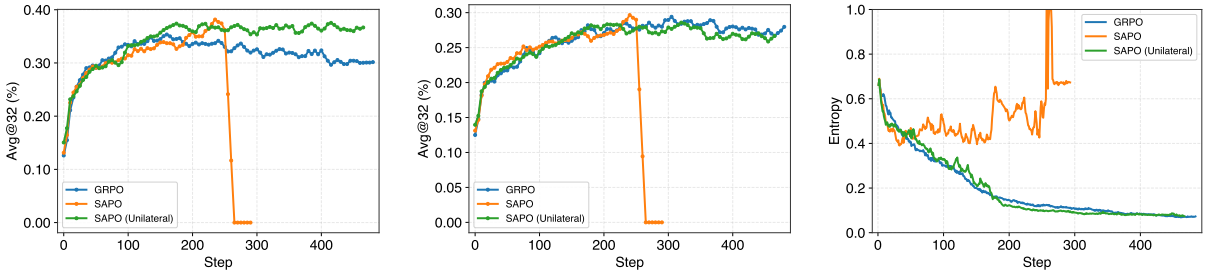
Figure 7 and Table 5 detail the full hyperparameter sweep on DeepSeek-R1-Distill-Qwen-7B. The results confirm that alongside the primary $\alpha = 0.3$, the $\alpha = 0.5$ also delivers robust and superiority.

D.4 Ablation on Positive Risk Control

We conducted a decoupling experiment on DeepSeek-R1-Distill-Qwen-1.5B with $\alpha = 0.5$ and $\beta_{\text{low}} = 0.03$, comparing $\beta_{\text{high}} = 0$ against $\beta_{\text{high}} = 0.03$. The training dynamics in Figure 8 demonstrate that without risk control on positive samples ($\beta_{\text{high}} = 0$), performance begins to degrade prematurely. Conversely, enabling positive risk control prevents early degeneration and promotes sustained exploration.

D.5 Scalability and Detailed Performance

To further investigate the scalability of our approach, we applied MASPO to the larger DeepSeek-R1-Distill-Qwen-14B model. As shown in Table 6, MASPO demonstrates significant scalability, consistently outperforming the GRPO baseline across all benchmarks with a +2.8% gain in average Avg@32. Furthermore, to provide a more fine-grained understanding of the generation quality, we report the detailed Pass@k performance on AIME 2024 and AIME 2025 in Table 7. These results confirm that MASPO not only achieves higher accuracy (Pass@1) but also maintains a higher upper bound of correct solutions (Pass@32) compared to other strong baselines.

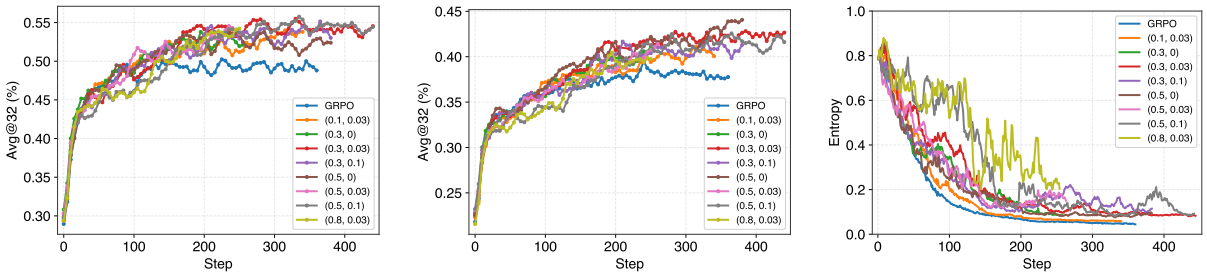


(a) AIME24 Accuracy (b) AIME25 Accuracy (c) Policy Entropy

Figure 6: Ablation study of the SAPO gating mechanism on the 1.5B model. The unilateral design (similar to MASPO) prevents the training collapse observed in the original bilateral SAPO.

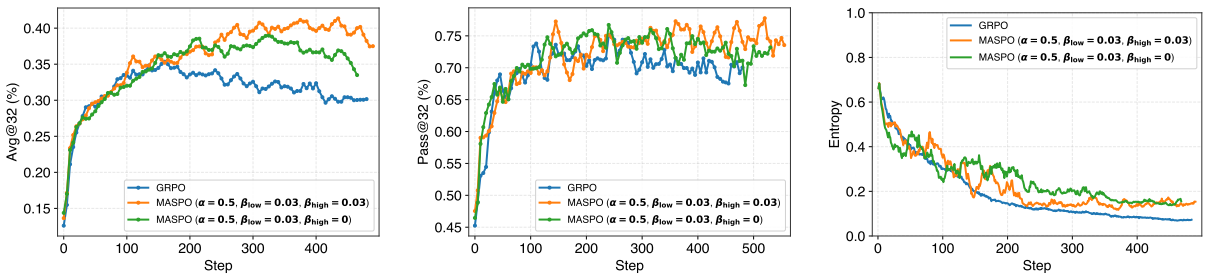
Table 5: Full hyperparameters comparison on DeepSeek-R1-Distill-Qwen-7B.

Hyperparameters	AIME24		AIME25		AMC23		MATH500		Minerva		Olympiad		Avg.	
	A@32	P@32	A@32	P@32	A@32	P@32	A@32	P@32	A@32	P@32	A@32	P@32	A@32	P@32
$\alpha = 0.1, \beta = 0.03$	53.1	78.4	41.3	72.0	89.5	97.5	84.1	90.8	38.1	55.5	56.2	70.6	60.4	77.5
$\alpha = 0.3, \beta = 0$	54.8	79.8	39.8	58.7	90.3	96.6	84.7	92.9	38.2	55.2	57.3	73.6	60.9	76.1
$\alpha = 0.3, \beta = 0.03$	53.2	82.4	**42.9**	**73.2**	91.4	95.0	**86.0**	**94.7**	39.3	58.6	58.0	**74.9**	61.8	**79.8**
$\alpha = 0.3, \beta = 0.1$	55.3	**85.6**	40.1	66.3	90.9	**98.7**	84.1	92.9	38.2	55.0	55.8	71.4	60.7	78.3
$\alpha = 0.5, \beta = 0$	51.5	81.8	41.6	64.6	90.5	97.2	84.4	91.6	38.8	55.9	56.5	71.6	60.6	77.1
$\alpha = 0.5, \beta = 0.03$	55.0	82.4	42.5	65.2	91.6	98.3	85.3	92.3	39.3	55.2	**58.1**	72.3	**62.0**	77.6
$\alpha = 0.5, \beta = 0.1$	**56.3**	85.2	39.7	60.8	91.0	95.0	83.1	93.5	37.8	**58.8**	54.8	71.8	60.5	77.5
$\alpha = 0.8, \beta = 0.03$	54.9	82.7	40.5	64.8	**92.0**	97.5	84.8	92.3	**40.1**	56.4	56.0	72.6	61.4	77.7



(a) Dynamics of AIME24 Accuracy (b) Dynamics of AIME25 Accuracy (c) Dynamics of Policy Entropy

Figure 7: Complete hyperparameter sweep results on the DeepSeek-R1-Distill-Qwen-7B, confirming the robustness of MASPO across a wider range of hyperparameter settings.



(a) Dynamics of AIME24 Avg@32 (b) Dynamics of AIME24 Pass@32 (c) Dynamics of Policy Entropy

Figure 8: Ablation study on positive risk control (β_{high}) using the 1.5B model.

Table 6: Performance results on the DeepSeek-R1-Distill-Qwen-14B model. We report Avg@32 and Pass@32.

	AIME24		AIME25		AMC23		MATH500		Minerva		Olympiad		Average	
Method	A@32	P@32	A@32	P@32	A@32	P@32	A@32	P@32	A@32	P@32	A@32	P@32	A@32	P@32
GRPO	56.6	82.9	40.5	67.8	92.2	97.4	66.5	70.4	22.6	35.3	43.2	50.7	53.6	67.4
MASPO	**63.5**	**86.1**	**45.3**	**72.4**	**93.4**	**99.9**	**66.8**	**73.3**	**24.6**	**41.4**	**45.0**	**53.7**	**56.4**	**71.1**

Table 7: Detailed Pass@k performance on AIME 2024 and AIME 2025 across all model scales.

Method	**Pass@1**	**Pass@2**	**Pass@4**	**Pass@8**	**Pass@16**	**Pass@32**
Backbone: DeepSeek-R1-Distill-Qwen-1.5B (AIME 2024)						
GRPO	33.2	43.2	53.2	61.6	67.9	71.8
Clip Higher	36.6	44.8	52.4	59.7	65.9	70.1
DAC 400	40.0	49.7	58.4	65.3	70.2	73.6
Entropy Adv	29.6	39.8	49.0	56.5	62.6	67.8
BAPO	38.3	47.6	55.3	62.1	67.9	72.3
SAPO	39.3	48.3	56.3	63.4	69.6	73.7
MASPO	**41.0**	**51.4**	**60.3**	**67.5**	**72.2**	**74.8**
Backbone: DeepSeek-R1-Distill-Qwen-1.5B (AIME 2025)						
GRPO	27.7	33.0	38.1	42.6	46.5	49.9
Clip Higher	**30.1**	**34.8**	**39.7**	45.0	50.8	55.8
DAC 400	28.7	34.1	39.0	43.6	48.0	51.2
Entropy Adv	23.9	29.0	32.9	37.0	41.1	44.2
BAPO	28.0	33.0	37.5	42.8	49.3	55.7
SAPO	28.0	32.1	36.0	40.8	45.7	49.9
MASPO	28.4	33.9	39.3	**45.1**	**51.6**	**58.0**
Backbone: DeepSeek-R1-Distill-Qwen-7B (AIME 2024)						
GRPO	48.2	58.0	66.6	73.4	78.6	**82.5**
Clip Higher	47.4	56.7	65.0	72.6	78.4	82.4
DAC 400	50.2	59.5	67.0	72.8	77.9	81.2
Entropy Adv	49.1	59.8	68.9	75.4	79.4	81.5
BAPO	47.1	56.5	64.9	71.7	76.8	80.3
SAPO	47.5	56.7	64.6	71.3	76.3	79.2
MASPO	**53.2**	**66.2**	**70.0**	**75.6**	**79.9**	82.4
Backbone: DeepSeek-R1-Distill-Qwen-7B (AIME 2025)						
GRPO	37.4	44.4	50.8	55.2	58.1	60.5
Clip Higher	37.4	44.9	52.1	58.2	63.3	67.2
DAC 400	36.2	43.4	50.6	56.3	59.6	61.7
Entropy Adv	34.4	41.0	46.4	50.3	54.0	58.1
BAPO	37.7	44.5	50.3	54.2	56.7	58.4
SAPO	35.3	41.4	47.4	51.6	55.3	58.0
MASPO	**42.9**	**50.7**	**57.3**	**63.4**	**68.9**	**73.2**
Backbone: DeepSeek-R1-Distill-Qwen-14B						
Dataset: AIME 2024						
GRPO	60.0	69.3	76.0	79.8	81.9	82.9
MASPO	**63.5**	**72.4**	**77.7**	**81.4**	**84.4**	**86.1**
Dataset: AIME 2025						
GRPO	42.0	48.4	54.2	60.1	64.7	67.8
MASPO	**45.3**	**52.2**	**58.2**	**63.7**	**68.4**	**72.4**