DISPO: Enhancing Training Efficiency and Stability in Reinforcement Learning for Large Language Model Mathematical Reasoning

Batuhan K. Karaman[1,*], **Aditya Rawal**[2], **Suhaila Shakiah**[2]
Mohammad Ghavamzadeh[3], **Arijit Biswas**[2], **Ruida Zhou**[2,†]
[1]Cornell University [2]Amazon AGI [3]Qualcomm AI Research
[*]Work done during an internship at Amazon AGI
[†]Corresponding author: zruida@amazon.com

Abstract

Reinforcement learning with verifiable rewards has emerged as a promising paradigm for enhancing the reasoning capabilities of large language models particularly in mathematics. Current approaches in this domain present a clear trade-off: PPO-style methods (e.g., GRPO/DAPO) offer training stability but exhibit slow learning trajectories due to their trust-region constraints on policy updates, while REINFORCE-style approaches (e.g., CISPO) demonstrate improved learning efficiency but suffer from performance instability as they clip importance sampling weights while still permitting non-zero gradients outside the trust-region. To address these limitations, we introduce DISPO, a simple yet effective REINFORCE-style algorithm that *decouples* the up-clipping and down-clipping of importance sampling weights for correct and incorrect responses, yielding four controllable policy update regimes. Through targeted ablations, we uncover how each regime impacts training: for correct responses, weights > 1 increase the average token entropy (i.e., exploration) while weights < 1 decrease it (i.e., distillation) – both beneficial but causing gradual performance degradation when excessive. For incorrect responses, overly restrictive clipping triggers sudden performance collapse through repetitive outputs (when weights > 1) or vanishing response lengths (when weights < 1). By separately tuning these four clipping parameters, DISPO maintains the exploration-distillation balance while preventing catastrophic failures, achieving 61.04% on AIME'24 (vs. 55.42% CISPO and 50.21% DAPO) with similar gains across various benchmarks and models.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

1 INTRODUCTION

Recent large language models (LLMs) such as DeepSeek-R1 [Guo et al., 2025], Qwen3 [Yang et al., 2025], OpenAI o1 [Jaech et al., 2024], and Claude Sonnet 3.5 [Anthropic, 2024] have demonstrated strong performance on reasoning tasks, including mathematical problem-solving, logical deduction, and scientific analysis. These tasks demand maintaining coherent chains of thought (CoT) while exploring diverse solution paths. A key driver of these advances is reinforcement learning with verifiable rewards (RLVR) [Yang et al., 2025, Guo et al., 2025, Lambert et al., 2024], where models optimize outputs through RL objectives directly tied to response correctness.

PPO-style RLVR algorithms such as GRPO [Shao et al., 2024] and DAPO [Yu et al., 2025] dominate large-scale deployments (e.g., DeepSeekMath, Qwen) due to their stability from trust-region-like constraints, though at the cost of slower learning. Recent work has revisited REINFORCE [Williams, 1992], which allows non-zero gradients outside the trust-region. REINFORCE-style algorithms such as CISPO [Chen et al., 2025] can surpass PPO methods with far fewer updates, offering appealing training efficiency [Ahmadian et al., 2024, Arnal et al., 2025]. However, this efficiency compromises stability: Zheng et al. [2025] reported sudden performance collapses in CISPO, while Arnal et al. [2025] documented unstable REINFORCE dynamics. Figure 1 illustrates this trade-

off—DAPO remains stable but slow (blue), while REINFORCE trains efficiently initially but collapses later (orange).

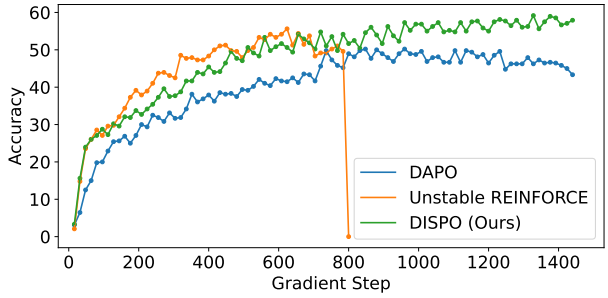

Figure 1: Learning curves of RLVR algorithms.

In this work, we introduce DISPO (**D**ecoupled **I**mportance **S**ampling-weighted **P**olicy **O**ptimization), a simple yet effective REINFORCE-style algorithm that achieves the best of both worlds: maintaining efficiency while ensuring stability that enables longer training without performance collapse, as shown in Figure 1 (green). DISPO assigns distinct upper and lower clipping bounds conditioned on (i) reward sign (correct vs. incorrect) and (ii) whether the importance sampling (IS) weight is above or below 1, yielding four controllable policy update regimes, as illustrated in Figure 2. Through targeted ablations, we isolate and reveal each regime's distinct impact on training dynamics. For correct responses, importance weights > 1 amplify token entropy to promote exploration (Regime 1), while weights < 1 suppress entropy for distillation (Regime 2)—that is, concentrating probability on the tokens leading to the correct response. Both effects are beneficial in moderation but can cause gradual performance degradation when excessive. For incorrect responses, overly restrictive clipping triggers catastrophic failures: repetitive outputs emerge when weights > 1 (Regime 3), while vanishingly short responses occur when weights < 1 (Regime 4). Our key findings are:

1. The clipping parameters for Regime 1 (exploration) and Regime 2 (distillation) have opposing effects on entropy and can be tuned jointly to control the exploration-distillation balance and mitigate gradual performance degradation.

2. Unlike the gradual degradation in correct responses, incorrect responses exhibit sudden collapses when clipping bounds are too restrictive: insufficient relaxation for Regime 3 causes repetitive outputs, while over-restriction for Regime 4 drives response lengths toward zero. For stable training, neither regime should be overly constrained.

3. By applying carefully tuned clipping bounds for each regime, DISPO balances exploration and distillation while preventing catastrophic failures. DISPO achieves 61.04% on AIME'24 (vs. 55.42% for CISPO and 50.21% for DAPO), with similar gains across various benchmarks and models.

Outline Section 2 reviews related work, while Section 3 provides necessary background to establish the foundation for DISPO. Section 4 introduces DISPO and presents our methodology for analyzing its four policy update regimes. Section 5 presents our experimental results, comparing DISPO against baselines and examining each policy update regime in detail.

2 RELATED WORK

Foundations of Policy Gradient Methods The REINFORCE algorithm [Williams, 1992] established the foundation for policy gradient methods in Reinforcement Learning (RL) by demonstrating gradient-based optimization for stochastic policies. Building on this, Proximal Policy Optimization (PPO) [Schulman et al., 2017] introduced a clipped surrogate objective to approximate a trust-region, improving stability. PPO has since become widely adopted across domains, such as robotics [Akkaya et al., 2019], game playing [Berner et al., 2019], and continuous control/locomotion [Heess et al., 2017].

RL for Language Models The application of RL to language models began with Reinforcement Learning from Human Feedback (RLHF) [Christiano et al., 2017, Stiennon et al., 2022, Ouyang et al., 2022, Lightman et al., 2023], which uses human preferences to align model outputs, later extended by RLAIF [Lee et al., 2024, Bai et al., 2022] using AI feedback instead. Recently, Reinforcement Learning with Verifiable Rewards (RLVR) has emerged for domains with automated verification, particularly mathematical reasoning [Uesato et al., 2022, Wang et al., 2023], eliminating human annotation requirements. Within RLVR, two algorithmic families have emerged: PPO-style methods (GRPO [Shao et al., 2024], DAPO [Yu et al., 2025] and others [Liu et al., 2025]) maintain stability but converge slowly, while REINFORCE-style approaches offer faster learning. CISPO [Chen et al., 2025] introduces clipped importance sampling to REINFORCE, Arnal et al. [2025] uses separate learning rates for positive/negative rewards, and Ahmadian et al. [2024] shows REINFORCE with careful tuning can match complex algorithms. However, these methods exhibit training instability [Zheng et al., 2025, Arnal et al., 2025], motivating our work.

$$J_{\text{REINFORCE}}(\theta) = \mathbb{E}_{(q,a)\sim\mathcal{D},\, o_i\sim\pi_{\text{ref}}(\cdot|q)}\left[\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\text{sg}\big(r_{i,t}(\theta)\big)\, A_{i,t}\, \log\pi_\theta(o_{i,t}\mid q, o_{i,<t})\right]$$

$$J_{\text{DISPO}}(\theta) = \mathbb{E}_{(q,a)\sim\mathcal{D},\, \{o_i\}_{i=1}^{G}\sim\pi_{\text{ref}}(\cdot|q)}\left[\frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}\text{sg}\big(r^d_{i,t}(\theta)\big)\, \hat{A}_{i,t}\, \log\pi_\theta(o_{i,t}\mid q, o_{i,<t})\right]$$

Controls Regime 2 (Provides distillation)

Controls Regime 1 (Provides exploration)

$$r^d_{i,t}(\theta) = \begin{cases}\text{clip}\big(r_{i,t}(\theta);\, 1-\epsilon^+_{\text{low}},\, 1+\epsilon^+_{\text{high}}\big), & \hat{A}_{i,t} > 0,\\ \text{clip}\big(r_{i,t}(\theta);\, 1-\epsilon^-_{\text{low}},\, 1+\epsilon^-_{\text{high}}\big), & \hat{A}_{i,t} < 0\end{cases}$$

Controls Regime 4 (Prevents zero response length)

Controls Regime 3 (Prevents falling into repetitions)

Figure 2: DISPO extends REINFORCE with (i) group-relative advantage estimation, (ii) token-level normalization, and (iii) decoupled IS weight $r^d_{i,t}(\theta)$. Each ϵ in the decoupled IS weight controls a distinct policy update regime.

3 BACKGROUND

We consider the RLVR setting for LLM reasoning, where π_θ denotes an LLM policy parameterized by θ that randomly predicts next token in the token space $\mathcal{T}$. Given a dataset $\mathcal{D}$ of question-answer pairs (q, a), where $q \in \mathcal{T}^*$, $a \in \mathcal{T}^*$, and $\mathcal{T}^*$ is the space of token sequences, the model generates a response o by sampling tokens autoregressively: $o = (o_1, o_2, \ldots, o_{|o|}) \in \mathcal{T}^{|o|}$, where $o_t \sim \pi_\theta(\cdot|q, o_{<t})$. The correctness of response o for question q is judged by a verifiable reward function $R(o, a) \in \{-1, 1\}$ based on the ground truth answer a.

GRPO Algorithm For each question q, GRPO samples G responses from a frozen reference snapshot of the model, denoted by π_{ref}. The objective adopts PPO's clipped surrogate as

$$J_{\text{GRPO}}(\theta) = \mathbb{E}_{(q,a)\sim\mathcal{D},\, \{o_i\}_{i=1}^{G}\sim\pi_{\text{ref}}(\cdot|q)}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\min\big(r_{i,t}(\theta)\hat{A}_{i,t},\ r^g_{i,t}(\theta)\hat{A}_{i,t}\big)\right], \quad (1)$$

where $r_{i,t}(\theta) = \frac{\pi_\theta(o_t|q,o_{<t})}{\pi_{\text{ref}}(o_t|q,o_{<t})}$ is the importance-sampling (IS) weight and

$$r^g_{i,t}(\theta) = \text{clip}\big(r_{i,t}(\theta);\, 1-\epsilon,\, 1+\epsilon\big) \quad (2)$$

is its clipped counterpart. The clip function is defined as $\text{clip}(x; a, b) := \min\big(\max(x, a),\, b\big)$. The group-relative advantage estimation for the i-th response (constant across t) is

$$\hat{A}_{i,t} = \frac{R_i - \mu_G}{\sigma_G}, \quad (3)$$

where $R_i \in \{-1, 1\}$ is the binary verifier reward, and $\mu_G = \frac{1}{G}\sum_{j=1}^{G} R_j$ and $\sigma_G = \sqrt{\frac{1}{G}\sum_{j=1}^{G}(R_j - \mu_G)^2}$ are the group mean and standard deviation. Thus, $\hat{A}_{i,t} > 0$ indicates a correct response, whereas $\hat{A}_{i,t} < 0$ indicates an incorrect one. Together, PPO-style clipping (Eq. 2) and the $\min(\cdot)$ surrogate bound effective updates when $r_{i,t}(\theta)$ leaves $[1-\epsilon, 1+\epsilon]$, preserve a trust-region–like stability. We note that we omit the KL-regularization term in Eq. 1 for brevity.

DAPO Algorithm DAPO [Yu et al., 2025] extends GRPO with (i) asymmetric clipping bounds $\epsilon_{\text{high}} > \epsilon_{\text{low}}$ to promote exploration, (ii) dynamic sampling that filters out uninformative groups (e.g., all-correct or all-incorrect) before the update, (iii) token-level normalization via $\frac{1}{\sum_{i=1}^{G}|o_i|}$, and (iv) an overlong penalty term that discourages excessively long responses. The DAPO objective is given as:

$$J_{\text{DAPO}}(\theta) = \mathbb{E}_{(q,a)\sim\mathcal{D},\, \{o_i\}_{i=1}^{G}\sim\pi_{\text{ref}}(\cdot|q)}\left[\frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}\min\big(r_{i,t}(\theta)\hat{A}_{i,t},\ r^c_{i,t}(\theta)\hat{A}_{i,t}\big)\right], \quad (4)$$

where the clipped IS weight $r^c_{i,t}(\theta)$ is defined as

$$r^c_{i,t}(\theta) = \text{clip}\big(r_{i,t}(\theta);\ 1-\epsilon_{\text{low}},\ 1+\epsilon_{\text{high}}\big). \quad (5)$$

Similar to GRPO, the asymmetric window $(\epsilon_{\text{low}}, \epsilon_{\text{high}})$ in DAPO preserves a trust-region–like stability.

REINFORCE Algorithm The objective of off-policy REINFORCE at the token-level can be written as

$$J_{\text{REINFORCE}}(\theta) = \mathbb{E}_{(q,a)\sim\mathcal{D},\, o_i\sim\pi_{\text{ref}}(\cdot|q)}\left[\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\text{sg}\big(r_{i,t}(\theta)\big)\, A_{i,t}\log\pi_\theta(o_{i,t}\mid q, o_{i,<t})\right], \quad (6)$$

where $A_{i,t}$ is the advantage and $\mathrm{sg}(\cdot)$ denotes the stop-gradient operator, i.e., the IS weight $r_{i,t}(\theta)$ still weights the loss but it is not differentiated. Unlike PPO/GRPO/DAPO, REINFORCE imposes no trust-region constraint; thus gradients can flow even when the IS weight $r_{i,t}(\theta)$ deviates substantially from 1.

CISPO Algorithm CISPO extends off-policy REINFORCE with: (i) group-sampling with group-size G (as in GRPO); and (ii–v) the DAPO-style components: asymmetric clipping bounds $(\epsilon_{\text{low}}, \epsilon_{\text{high}})$, dynamic sampling, token-level normalization, and an overlong penalty term. The CISPO objective is

$$J_{\text{CISPO}}(\theta) = \mathbb{E}_{(q,a)\sim\mathcal{D},\, \{o_i\}_{i=1}^{G}\sim\pi_{\text{ref}}(\cdot|q)} \left[\frac{1}{\sum_{i=1}^{G}|o_i|} \sum_{i=1}^{G}\sum_{t=1}^{|o_i|} \mathrm{sg}\big(r^{c}_{i,t}(\theta)\big)\, \hat{A}_{i,t}\, \log \pi_\theta(o_{i,t} \mid q, o_{i,<t}) \right], \tag{7}$$

where the clipped IS weight $r^{c}_{i,t}(\theta)$ is defined as in Eq. 5. Similar to REINFORCE, CISPO imposes no trust-region constraint: gradients can flow for every token, even though their effect is clipped.

Notably, CISPO applies the same asymmetric clipping window to tokens from both correct and incorrect responses. We show that this uniform treatment overlooks the fundamentally different optimization dynamics across REINFORCE's four policy update regimes, contributing to training instability and limited exploration—motivating our proposed decoupled clipping strategy in DISPO.

4 DECOUPLED IMPORTANCE SAMPLING-WEIGHTED POLICY OPTIMIZATION (DISPO)

DISPO extends REINFORCE by using group-relative advantage estimation and token-level normalization, similar to CISPO. Besides, we introduce separate clipping bounds for the IS weights in correct and incorrect responses. Formally, the DISPO objective is defined as

$$J_{\text{DISPO}}(\theta) = \mathbb{E}_{(q,a)\sim\mathcal{D},\, \{o_i\}_{i=1}^{G}\sim\pi_{\text{ref}}(\cdot|q)} \left[\frac{1}{\sum_{i=1}^{G}|o_i|} \sum_{i=1}^{G}\sum_{t=1}^{|o_i|} \mathrm{sg}\big(r^{d}_{i,t}(\theta)\big)\, \hat{A}_{i,t}\, \log \pi_\theta(o_{i,t} \mid q, o_{i,<t}) \right], \tag{8}$$

where the decoupled IS weight $r^{d}_{i,t}(\theta)$ is given by

$$r^{d}_{i,t}(\theta) = \begin{cases} \mathrm{clip}\big(r_{i,t}(\theta);\, 1-\epsilon^{+}_{\text{low}},\, 1+\epsilon^{+}_{\text{high}}\big), & \hat{A}_{i,t} > 0, \\ \mathrm{clip}\big(r_{i,t}(\theta);\, 1-\epsilon^{-}_{\text{low}},\, 1+\epsilon^{-}_{\text{high}}\big), & \hat{A}_{i,t} < 0. \end{cases} \tag{9}$$

This decoupled clipping strategy provides fine-grained control over the four distinct policy update regimes in REINFORCE, allowing us to amplify/suppress gradients for both correct and incorrect responses. In Section 4.1, we decompose these regimes systematically and present our ablation methodology to isolate their individual effects on training dynamics. We note that, following DAPO and CISPO, DISPO also incorporates dynamic sampling and an overlong penalty term in the loss calculation. A concise and demonstrative implementation of DISPO can be found at `https://github.com/batuhankmkaraman/dispo`.

Gradient-weight view We visualize how different objectives modulate the *magnitude* of the policy-gradient as a function of the importance ratio $r_{i,t}(\theta)$. For all the policy objectives discussed above, the gradient can be written as a function proportional to $\mathrm{sg}(w_{i,t}(\theta)\, r_{i,t}(\theta))\, \nabla_\theta \log \pi_\theta(y_{i,t} \mid q, y_{i,<t})\, \hat{A}_{i,t}$ ignoring the length normalization terms. The *gradient weight* $w_{i,t}(\theta)$ captures the algorithm-specific gating/clipping effect and should be interpreted as a relative scaling factor (not the full gradient), showing how each method attenuates or preserves the update when $r_{i,t}(\theta)$ deviates from 1. Figure 3 provides an intuitive view of DAPO, CISPO, and DISPO by showing their gradient-weight profiles as functions of the importance-sampling weight. We observe that PPO-style clipping (DAPO) enforces a hard cutoff, setting the update weight to zero once $r_{i,t}(\theta)$ leaves the trust region, whereas REINFORCE-style variants act more like *soft gates*. DISPO uses sign-dependent ratio control, applying different gating profiles for $\hat{A}_{i,t} > 0$ and $\hat{A}_{i,t} < 0$, which yields asymmetric gradient weighting as a function of $r_{i,t}(\theta)$.

4.1 Analyzing the Policy Update Regimes

In off-policy training, the IS weight $r_{i,t}(\theta)$ is inherited from the policy updated in the previous step and influences the gradient at the current step. At the very first update, $r_{i,t}(\theta) = 1$, so the contribution of each token to the gradient of the DISPO objective in Eq. 8 takes the form

$$\nabla_\theta J \propto \hat{A}_{i,t} \nabla_\theta \log \pi_\theta(o_{i,t} \mid q, o_{i,<t}). \tag{10}$$

However, as training progresses, $r_{i,t}(\theta)$ may drift above or below 1, reflecting how the current policy π_θ deviates from the reference policy π_{ref}.

We decompose the policy update regimes in DISPO along two axes: (1) whether the response is correct ($\hat{A}_{i,t} > 0$) or incorrect ($\hat{A}_{i,t} < 0$), and (2) whether the IS weight amplifies ($r_{i,t}(\theta) > 1$) or suppresses ($r_{i,t}(\theta) < 1$) the gradient. This yields four distinct update regimes that we will describe in detail below:

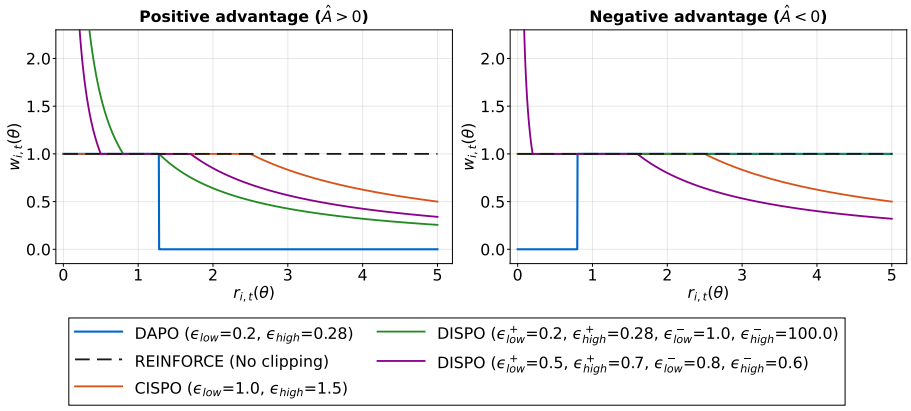

Figure 3: Gradient weight $w_{i,t}(\theta)$ as a function of the importance-sampling weight $r_{i,t}(\theta)$.

Regime 1: Amplified Positive Updates ($\hat{A}_{i,t} > 0$, $r_{i,t}(\theta) > 1$) This regime captures tokens in correct responses whose probabilities have increased relative to the reference policy during previous updates, as illustrated in Figure 4. We can write the gradient contribution of these tokens as

$$\nabla_\theta J \propto \mathrm{sg}\left(r^d_{i,t}(\theta)\right)\hat{A}_{i,t}\nabla_\theta \log \pi_\theta(o_{i,t} \mid q, o_{i,<t}). \quad (11)$$

With $r_{i,t}(\theta) > 1$, this regime amplifies the positive learning signal beyond the baseline gradient in Eq. 10, reinforcing the tokens that the model has already learned to favor. We note that increasing ϵ^+_{high} in Eq. 9 allows greater amplification by permitting larger values of $r^d_{i,t}(\theta)$, while setting $\epsilon^+_{\text{high}} = 0$ disables Regime 1 entirely. We observe in our experiments that Regime 1 increases average token-level entropy, serving as a key driver of exploration during training (Section 5.3).

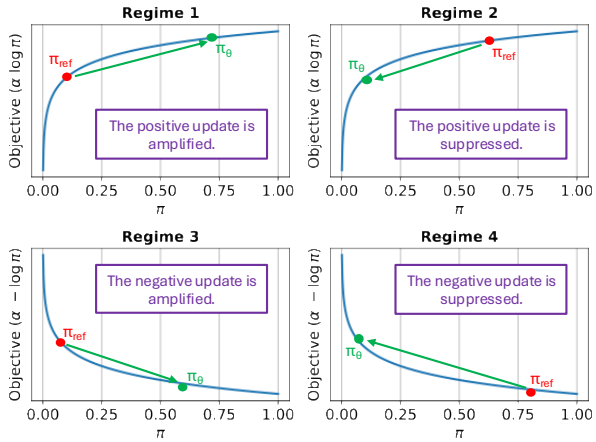

Figure 4: DISPO's four policy update regimes.

Regime 2: Suppressed Positive Updates ($\hat{A}_{i,t} > 0$, $r_{i,t}(\theta) < 1$) Here, tokens in correct responses have decreased in probability relative to the reference policy, as shown in Figure 4. The gradient expression is identical to Eq. 11, but since $r_{i,t}(\theta) < 1$, the positive update signal is suppressed. We note that increasing ϵ^+_{low} in Eq. 9 results in greater suppression by allowing smaller values of $r^d_{i,t}(\theta)$, while setting $\epsilon^+_{\text{low}} = 0$ disables Regime 2 entirely. We observe that Regime 2 reduces average token-level entropy, acting as a distillation mechanism that consolidates learned patterns during training (Section 5.3).

Regime 3: Amplified Negative Updates ($\hat{A}_{i,t} < 0$, $r_{i,t}(\theta) > 1$) This regime captures tokens in incorrect responses whose probabilities have increased relative to the reference policy during previous updates, as displayed in Figure 4. The gradient contribution of the tokens becomes

$$\nabla_\theta J \propto \mathrm{sg}\left(r^d_{i,t}(\theta)\right)\hat{A}_{i,t}\nabla_\theta \log \pi_\theta(o_{i,t} \mid q, o_{i,<t}). \quad (12)$$

With $\hat{A}_{i,t} < 0$ and $r_{i,t}(\theta) > 1$, this regime amplifies the negative learning signal, driving stronger unlearning of the tokens that the model has erroneously learned to favor. We note that increasing ϵ^-_{high} allows greater amplification by permitting larger values of $r^d_{i,t}(\theta)$, and setting $\epsilon^-_{\text{high}} = 0$ disables Regime 3 entirely. We observe that insufficient amplification in Regime 3 leads to repetition-induced collapse, where the model fails to adequately unlearn erroneous patterns (Section 5.3).

Regime 4: Suppressed Negative Updates ($\hat{A}_{i,t} < 0$, $r_{i,t}(\theta) < 1$) Finally, tokens in incorrect responses whose probabilities have decreased relative to the reference policy fall into this regime, as illustrated in Figure 4. As in Eq. 12, the negative advantage pushes the model to further reduce these probabilities, but since $r_{i,t}(\theta) < 1$, the unlearning signal is dampened. We note that increasing ϵ^-_{low} results in greater suppression by allowing smaller values of $r^d_{i,t}(\theta)$, while setting $\epsilon^-_{\text{low}} = 0$ disables Regime 4 entirely. We observe that excessive suppression in Regime 4 causes response

lengths to approach zero, indicating over-aggressive unlearning that disrupts the generation capability of the model (Section 5.3).

To isolate each regime's effect on training dynamics, we design controlled ablations by varying the four clipping parameters in Eq. 9. Setting any parameter to zero disables its corresponding regime, while positive values control the strength of amplification or suppression. Table 1 presents our ablation configurations. For Regimes 1 and 2, our baseline sets $\epsilon^{+}_{\text{low}} = \epsilon^{+}_{\text{high}} = 0$ and zeros gradients for incorrect responses, equivalent to online supervised fine-tuning (SFT). For Regimes 3 and 4, our baseline is DISPO with both regimes enabled, as both are necessary for stable training. To show their necessity, we disable each regime individually by setting $\epsilon^{-}_{\text{high}} = 0$ (Regime 3) or $\epsilon^{-}_{\text{low}} = 0$ (Regime 4).

5 RESULTS AND DISCUSSION

We first briefly describe our experimental setup, followed by the main results comparing DISPO against baseline methods. We then examine each policy update regime in detail, highlighting insights that also informed the design of DISPO.

5.1 Experimental Setup

We evaluate DISPO against PPO-style (DAPO) and REINFORCE-style (CISPO) baselines across diverse model sizes and architectures: Qwen3-8B-Base and Qwen3-14B-Base (both dense models), and Qwen3-30B-A3B-Base (MoE with 3.3B activated parameters). All models are trained on DAPO-Math-17k dataset [Yu et al., 2025, Hendrycks et al., 2021b, Saxton et al., 2019, Cobbe et al., 2021], and evaluated on five mathematical reasoning benchmarks: AIME'24 [MAA, 2025], AIME'25 [MAA, 2025], AMC'23 [MAA, 2024], MATH-500 [Hendrycks et al., 2021a], and Minerva [Lewkowycz et al., 2022]. We use Qwen3-14B-Base for our ablation studies on policy update regimes. Moreover, we use the advantage formulation in Eq. 3 in all experiments, including the ablations for Regimes 1 and 2. Additional information about the baseline methods, training, and evaluation can be found in Appendix B.

5.2 DISPO vs. SOTA Methods

DISPO outperforms the baselines significantly. Table 2 presents the evaluation of DISPO against the baselines across multiple mathematical reasoning benchmarks for all tested models. On the AIME'24 dataset, DISPO achieves substantial improvements: 61.04% accuracy on Qwen3-14B compared to 50.21% for DAPO and 55.42% for CISPO, representing a 10.83 percentage point improvement over DAPO. Similar patterns hold across other benchmarks, with DISPO showing particularly strong gains on competition-level problems (AIME'25: 45.83% vs. 38.96% for DAPO; AMC'23: 92.03% vs. 87.66% for DAPO). The improvements remain consistent across different model sizes and architectures—from the 8B dense model to the 30B mixture-of-experts (MoE) variant.

DISPO balances exploration and distillation while maintaining training stability. Figure 5 shows the learning curves of AIME'24 for DISPO and baseline methods. The entropy curves (bottom panel) reveal clear differences. We observe that CISPO loses entropy throughout training, whereas DISPO exhibits an increase. This difference is significant, as higher entropy indicates greater token-level exploration [Wang et al., 2025], and it explains DISPO's superior performance. DAPO also exhibits a similar rise in entropy, but its performance is significantly lower because it relies on token-clipped PPO rather than REINFORCE. Specifically, it discards tokens with low reference likelihood that serve as key entropy drivers—such as "but", "aha", and "since"—whose importance has been highlighted in prior work [Chen et al., 2025, Wang et al., 2025]. Learning curves for additional models (Qwen3-30B-A3B-Base and Qwen3-8B-Base) are shown in Appendix C, and they show similar patterns. Response length curve for Qwen3-14B-Base can also be found in Appendix C.

Regarding clipping bounds, we adopt the default values for DAPO: $\epsilon_{\text{low}} = 0.2$ and $\epsilon_{\text{high}} = 0.28$. Since the original CISPO clipping bounds were not released, we implemented it with $\epsilon_{\text{low}} = 1$ and $\epsilon_{\text{high}} = 100$, which proved stable across all evaluated models. Using either $\epsilon_{\text{low}} < 1$ or $\epsilon_{\text{high}} < 100$ in CISPO caused sudden performance collapse. Note that $\epsilon = 100$ effectively disables clipping while still preventing infinite importance weights. For DISPO, we set $\epsilon^{+}_{\text{low}} = 0.2$, $\epsilon^{+}_{\text{high}} = 10$, $\epsilon^{-}_{\text{low}} = 1$, and $\epsilon^{-}_{\text{high}} = 100$. In Section 5.3, we examine how these four ϵ parameters shape training dynamics through their corresponding policy update regimes, explain the rationale behind our DISPO hyperparameter choices, and provide insights into the previously reported CISPO collapses [Zheng et al., 2025].

5.3 Analysis of Policy Update Regimes in Off-Policy REINFORCE

Increasing $\epsilon^{+}_{\text{high}}$ in Regime 1 improves exploration. Figure 6 shows the impact of allowing $r_{i,t} > 1$ through different $\epsilon^{+}_{\text{high}}$ values while maintaining $\epsilon^{+}_{\text{low}} = 0$. The online SFT baseline (blue) shows stable but slow accuracy improvement, while both mod-

Table 1: Ablation configurations for analyzing update regimes in off-policy REINFORCE. For Regimes 1-2, we start from an online SFT baseline and enable each regime individually. For Regimes 3-4, we start from the full DISPO configuration and disable each regime individually.

Configuration	Response Type	Clipping Parameters	Active Regimes
Starting from online SFT baseline (analysis of Regimes 1 and 2)			
Online SFT Baseline	Correct only	$\epsilon^+_{\text{low}} = 0$, $\epsilon^+_{\text{high}} = 0$	None ($\hat{r}_{i,t} = 1$)
+Regime 1	Correct only	$\epsilon^+_{\text{low}} = 0$, $\epsilon^+_{\text{high}} = 0.28$	Amplified Positive
+Regime 1	Correct only	$\epsilon^+_{\text{low}} = 0$, $\epsilon^+_{\text{high}} = 10$	Amplified Positive
+Regime 2	Correct only	$\epsilon^+_{\text{low}} = 0.2$, $\epsilon^+_{\text{high}} = 0$	Suppressed Positive
+Regime 2	Correct only	$\epsilon^+_{\text{low}} = 1$, $\epsilon^+_{\text{high}} = 0$	Suppressed Positive
Starting from DISPO baseline (analysis of Regimes 3 and 4)			
DISPO (Full)	Correct + Incorrect	$\epsilon^-_{\text{low}} = 1$, $\epsilon^-_{\text{high}} = 100$	All regimes
-Regime 3	Correct + Incorrect	$\epsilon^-_{\text{low}} = 1$, $\epsilon^-_{\text{high}} = 0$	without Amplified Negative
-Regime 4	Correct + Incorrect	$\epsilon^-_{\text{low}} = 0$, $\epsilon^-_{\text{high}} = 100$	without Suppressed Negative

Table 2: Comparison of DISPO with DAPO and CISPO across different model sizes and architectures. "–" denotes the performance of the starting checkpoint. The maximum values for each model and benchmark are highlighted in **bold**.

Benchmark (**Avg@16**)	RL Algorithm			
	–	DAPO	CISPO	DISPO
Model: Qwen3-14B-Base				
AIME'24	3.96	50.21	55.42	**61.04**
AIME'25	1.67	38.96	40.83	**45.83**
AMC'23	22.19	87.66	89.84	**92.03**
MATH-500	73.91	91.89	93.15	**94.61**
Minerva	36.10	45.22	45.66	**46.78**
Model: Qwen3-30B-A3B-Base (MoE)				
AIME'24	0.83	45.83	48.54	**53.75**
AIME'25	1.04	31.25	35.62	**38.12**
AMC'23	12.50	83.75	87.03	**89.38**
MATH-500	73.54	90.56	91.18	**92.59**
Minerva	30.35	43.45	44.26	**44.60**
Model: Qwen3-8B-Base				
AIME'24	4.58	38.33	43.54	**45.00**
AIME'25	2.71	27.50	30.83	**31.46**
AMC'23	24.69	80.94	80.94	**87.66**
MATH-500	71.22	91.25	90.74	**92.34**
Minerva	30.95	**44.85**	43.52	44.60

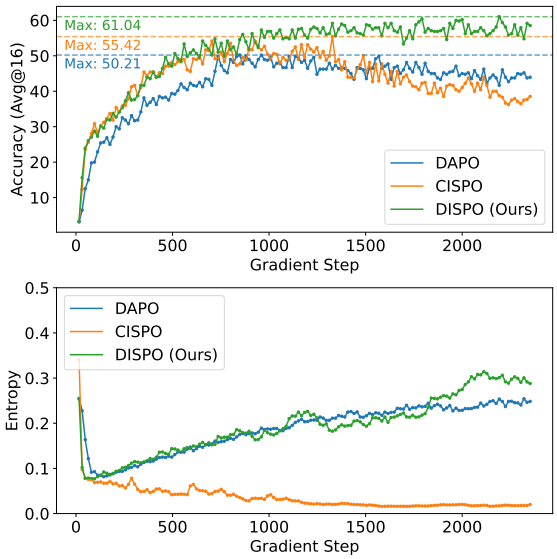

Figure 5: Accuracy and entropy curves of DAPO, CISPO, and DISPO.

els with $\epsilon^+_{\text{high}} > 0$ demonstrate faster gains, indicating improved training efficiency. The entropy curves reveal the underlying mechanism: the baseline maintains constant entropy throughout training, whereas configurations with $\epsilon^+_{\text{high}} > 0$ exhibit increasing entropy, reflecting progressive exploration. As we discussed in Section 5.2, amplifying gradient for low-reference-likelihood tokens in correct responses leads to more effective token utilization and increased exploration.

Excessive exploration causes gradual performance degradation. The increased exploration eventually becomes detrimental in later training stages. Both models with $\epsilon^+_{\text{high}} > 0$ show accuracy degradation after peaking, as excessive exploration leads to sampling increasingly unlikely tokens that harm reasoning coherence. This reveals a critical trade-off: amplification enhances learning efficiency through beneficial exploration, but excessive amplification causes uncontrolled exploration and gradual performance decline later in training. The model with $\epsilon^+_{\text{high}} = 0.28$ provides more controlled exploration than $\epsilon^+_{\text{high}} = 10$—entropy increases more gradually, enabling longer training before degradation and higher peak accuracy. This demonstrates that carefully tuning ϵ^+_{high} preserves efficiency benefits while mitigating uncon-

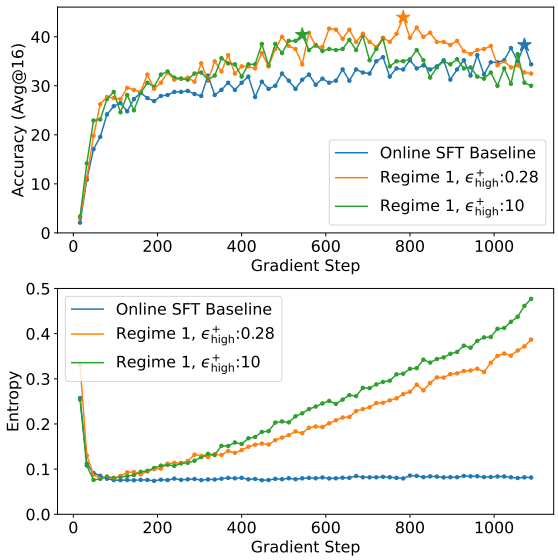

Figure 6: Accuracy and entropy curves of Regime 1 runs. ⋆ marks the maximum value.

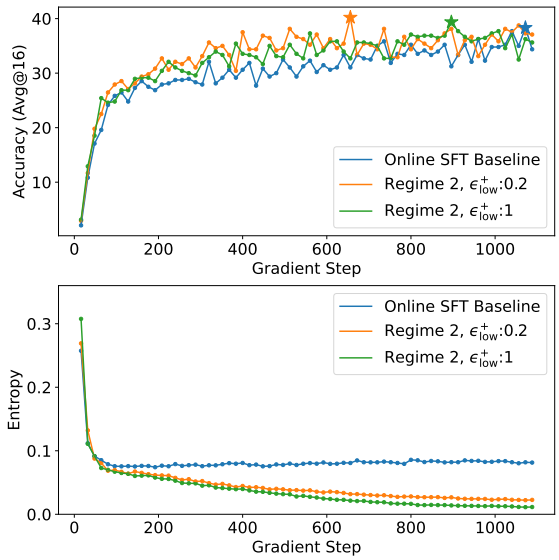

Figure 7: Accuracy and entropy curves of Regime 2 runs. ⋆ marks the maximum value.

trolled exploration risks. We note that DISPO in Figure 5 might also experience accuracy decline with extended training due to Regime 1, though computational constraints prevented us form verifying it. Importantly, in Regime 1, instability manifests itself as gradual degradation rather than sudden collapse.

Increasing ϵ^+_{low} in Regime 2 improves distillation. Figure 7 shows the impact of varying ϵ^+_{low} while maintaining $\epsilon_{\text{high}} = 0$. We see that allowing $r_{i,t} < 1$ yields efficiency gains compared to the online SFT baseline. As shown in Figure 7, Regime 2 decreases entropy—opposite to Regime 1's exploratory behavior—indicating the model actively prunes its token vocabulary. This entropy reduction reflects a distillation mechanism: by attenuating the learning signal for tokens with decreased probabilities ($r_{i,t} < 1$), the model filters out less reliable tokens even within correct responses. While some suppressed tokens may contain useful patterns, they likely include noise or suboptimal solution paths. This selective reinforcement accelerates learning by focusing on high-probability tokens that form the core problem-solving strategy, rather than indiscriminately reinforcing all tokens in correct solutions. Comparing $\epsilon^-_{\text{low}} = 0.2$ versus 1.0, the former enables more controlled distillation through gradual entropy reduction, achieving slightly higher peak accuracy.

Distillation offers limited standalone benefits but effectively counterbalances excessive exploration. While Regime 2 achieves efficiency gains (Figure 7), these improvements remain modest compared to Regime 1's exploration-driven performance (Figure 6). The decreasing entropy in Regime 2 signals premature convergence to limited solution strategies, as the model trades exploratory capacity for consistency and potentially misses innovative approaches requiring lower-probability tokens. This explains why pure distillation yields lower peak accuracy—exploration proves more crucial than consolidation for superior reasoning performance. Response length curves for both regimes appear in Appendix C. Notably, when Regimes 1 and 2 operate simultaneously (Appendix D), their competing entropy effects create a balanced exploration-distillation dynamic that achieves higher accuracy than either regime can achieve alone, demonstrating that distillation's primary value lies in moderating exploration rather than serving as a standalone strategy.

In Regime 3, setting $\epsilon^-_{\text{high}} > 0$ is necessary to prevent repetition collapse. Figure 8 shows that disabling Regime 3 ($\epsilon^-_{\text{high}} = 0$, orange) causes sudden accuracy collapse and response length spike early in training, compared to full DISPO (blue). Regime 3 targets incorrect tokens whose probabilities exceed the reference policy ($r_{i,t} > 1$). Without Regime 3's amplification, these tokens receive insufficient negative gradients for unlearning—since $|\nabla_\theta \log \pi_\theta|$ naturally decreases as $\pi_\theta \to 1$, high-probability tokens already have weak gradients. In essence, the model loses the ability to "forget" its mistakes, as the unlearning signal becomes too weak to overcome the reinforcement from previous updates. Setting $\epsilon^-_{\text{high}} = 0$ caps $r_{i,t}$ at 1, further weakens these gradients and prevents ade-

quate penalization. Consequently, the model repeatedly generates these incorrect high-probability tokens, causing the repetition-driven length spike in Figure 8 (bottom). Figure 13 in Appendix C shows a representative example. We hypothesize that this is one of the reasons behind previously reported CISPO collapses [Zheng et al., 2025]—our experiments confirm that $\epsilon^{-}_{\text{high}} < 1$ causes CISPO to fail with similar repetition patterns in our experimental setup.

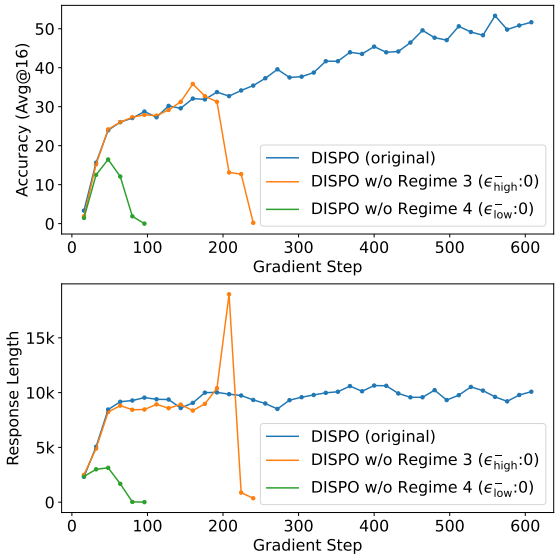

Figure 8: Accuracy and response length curves of runs in which Regimes 3 and 4 are disabled.

In Regime 4, setting $\epsilon^{-}_{\text{low}} > 0$ prevents length collapse from excessive unlearning. Figure 8 shows rapid deterioration in both accuracy and response length when Regime 4 is disabled ($\epsilon^{-}_{\text{low}} = 0$, green). Regime 4 governs incorrect tokens with probabilities below the reference policy ($r_{i,t} < 1$). These low-probability tokens naturally receive strong negative gradients since $|\nabla_\theta \log \pi_\theta|$ grows as $\pi_\theta \to 0$. Setting $\epsilon^{-}_{\text{low}} = 0$ removes suppression—allowing $r_{i,t}$ below 1 amplifies these already-strong gradients, causing excessive penalization that drives response length toward zero (Figure 8, bottom). This over-penalization essentially teaches the model to "give up" on generation rather than learn correct patterns. Thus, Regime 4's gradient suppression prevents over-correction on already-unlikely tokens. We hypothesize this is another reason behind CISPO collapses [Zheng et al., 2025], as our experiments show $\epsilon^{-}_{\text{low}} < 100$ causes similar length collapse in CISPO.

6 CONCLUSION

We presented DISPO, a simple yet effective modification to off-policy REINFORCE that decouples the clipping of importance sampling weights for correct and incorrect responses. Through systematic ablations of the four resulting policy update regimes, we identified distinct failure modes in off-policy REINFORCE-style methods. Our results demonstrate that properly balancing these regimes promotes exploration-distillation balance while maintaining training stability, achieving superior performance over existing methods.

References

A. Ahmadian, C. Cremer, M. Gallé, M. Fadaee, J. Kreutzer, O. Pietquin, A. Üstün, and S. Hooker. Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms, 2024. URL https://arxiv.org/abs/2402.14740.

I. Akkaya, M. Andrychowicz, M. Chociej, M. Litwin, B. McGrew, A. Petron, A. Paino, M. Plappert, G. Powell, R. Ribas, J. Schneider, N. Tezak, J. Tworek, P. Welinder, L. Weng, Q. Yuan, W. Zaremba, and L. Zhang. Solving rubik's cube with a robot hand, 2019. URL https://arxiv.org/abs/1910.07113.

Anthropic. Claude 3.7 sonnet, 2024. URL https://www.anthropic.com/claude/sonnet.

C. Arnal, G. Narozniak, V. Cabannes, Y. Tang, J. Kempe, and R. Munos. Asymmetric reinforce for off-policy reinforcement learning: Balancing positive and negative rewards, 2025. URL https://arxiv.org/abs/2506.20520.

Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. Mckinnon, C. Chen, C. Olsson, C. Olah, D. Hernandez, D. Drain, D. Ganguli, D. Li, E. Tran-Johnson, E. Perez, J. Kerr, J. Mueller, J. Ladish, J. Landau, K. Ndousse, K. Lukosuite, L. Lovitt, M. Sellitto, N. Elhage, N. Schiefer, N. Mercado, N. Dassarma, R. Lasenby, R. Larson, S. Ringer, S. Johnston, S. Kravec, E. Showk, S. Fort, T. Lanham, T. Telleen-Lawton, T. Conerly, T. Henighan, T. Hume, S. Bowman, Z. Hatfield-Dodds, B. Mann, D. Amodei, N. Joseph, S. Mccandlish, T. Brown, and J. Kaplan. Constitutional ai: Harmlessness from ai feedback, 2022. URL https://arxiv.org/pdf/2212.08073.

C. Berner, G. Brockman, B. Chan, V. Cheung, P. Dębiak, C. Dennison, D. Farhi, Q. Fischer, S. Hashme, C. Hesse, R. Józefowicz, S. Gray, C. Olsson, J. Pachocki, M. Petrov, H. P. d. O. Pinto,

J. Raiman, T. Salimans, J. Schlatter, J. Schneider, S. Sidor, I. Sutskever, J. Tang, F. Wolski, and S. Zhang. Dota 2 with large scale deep reinforcement learning. *arXiv:1912.06680 [cs, stat]*, 12 2019. URL https://arxiv.org/abs/1912.06680.

A. Chen, A. Li, B. Gong, B. Jiang, B. Fei, B. Yang, B. Shan, C. Yu, C. Wang, C. Zhu, C. Xiao, C. Du, C. Zhang, C. Qiao, C. Zhang, C. Du, C. Guo, D. Chen, D. Ding, D. Sun, D. Li, E. Jiao, H. Zhou, H. Zhang, H. Ding, H. Sun, H. Feng, H. Cai, H. Zhu, J. Sun, J. Zhuang, J. Cai, J. Song, J. Zhu, J. Li, J. Tian, J. Liu, J. Xu, J. Yan, J. Liu, J. He, K. Feng, K. Yang, K. Xiao, L. Han, L. Wang, L. Yu, L. Feng, L. Li, L. Zheng, L. Du, L. Yang, L. Zeng, M. Yu, M. Tao, M. Chi, M. Zhang, M. Lin, N. Hu, N. Di, P. Gao, P. Li, P. Zhao, Q. Ren, Q. Xu, Q. Li, Q. Wang, R. Tian, R. Leng, S. Chen, S. Chen, S. Shi, S. Weng, S. Guan, S. Yu, S. Li, S. Zhu, T. Li, T. Cai, T. Liang, W. Cheng, W. Kong, W. Li, X. Chen, X. Song, X. Luo, X. Su, X. Li, X. Han, X. Hou, X. Lu, X. Zou, X. Shen, Y. Gong, Y. Ma, Y. Wang, Y. Shi, Y. Zhong, Y. Duan, Y. Fu, Y. Hu, Y. Gao, Y. Fan, Y. Yang, Y. Li, Y. Hu, Y. Huang, Y. Li, Y. Xu, Y. Mao, Y. Shi, Y. Wenren, Z. Li, Z. Li, Z. Tian, Z. Zhu, Z. Fan, Z. Wu, Z. Xu, Z. Yu, Z. Lyu, Z. Jiang, Z. Gao, Z. Wu, Z. Song, and Z. Sun. Minimax-m1: Scaling test-time compute efficiently with lightning attention, 2025. URL https://arxiv.org/abs/2506.13585.

P. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences, 2017. URL https://arxiv.org/abs/1706.03741.

K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse, and J. Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, X. Zhang, X. Yu, Y. Wu, W. Z. F, Z. Gou, Z. Shao, Z. Li, Z. Gao, A. Liu, B. Xue, B. Wang, B. Wu, B. Feng, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan, D. Dai, D. Chen, D. Ji, E. Li, F. Lin, F. Dai, F. Luo, G. Hao, G. Chen, G. Li, H. Zhang, H. Bao, H. Xu, H. Wang, H. Ding, H. Xin, H. Gao, H. Qu, H. Li, J. Guo, J. Li, J. Wang, J. Chen, J. Yuan, J. Qiu, J. Li, C. J. L, J. Ni, J. Liang, J. Chen, K. Dong, K. Hu, K. Gao, K. Guan, K. Huang, K. Yu, L. Wang, L. Zhang, L. Zhao, L. Wang, L. Zhang, L. Xu, L. Xia, M. Zhang, M. Zhang, M. Tang, M. Li, M. Wang, M. Li, N. Tian, P. Huang, P. Zhang, Q. Wang, Q. Chen, Q. Du, R. Ge, R. Zhang, R. Pan, R. Wang, C. R. J, J. R. L, R. Chen, S. Lu, S. Zhou, S. Chen, S. Ye, S. Wang, S. Yu, S. Zhou, S. Pan, L. S. S, S. Zhou, S. Wu, S. Ye, T. Yun, T. Pei, T. Sun, T. Wang, W. Zeng, W. Zhao, W. Liu, W. Liang, W. Gao, W. Yu, W. Zhang, X. W. L, W. An, X. Liu, X. Wang, X. Chen, X. Nie, X. Cheng, X. Liu, X. Xie, X. Liu, X. Yang, X. Li, X. Su, X. Lin, L. X. Q, X. Jin, X. Shen, X. Chen, X. Sun, X. Wang, X. Song, X. Zhou, X. Wang, X. Shan, L. Y. K, W. Y. Q, W. Y. X, Y. Zhang, Y. Xu, Y. Li, Y. Zhao, Y. Sun, Y. Wang, Y. Yu, Y. Zhang, Y. Shi, Y. Xiong, Y. He, Y. Piao, Y. Wang, Y. Tan, Y. Ma, Y. Liu, Y. Guo, Y. Ou, Y. Wang, Y. Gong, Y. Zou, Y. He, Y. Xiong, Y. Luo, Y. You, Y. Liu, Y. Zhou, Z. Y. X, Y. Xu, Y. Huang, Y. Li, Y. Zheng, Y. Zhu, Y. Ma, Y. Tang, Y. Zha, Y. Yan, R. Z. Z, Z. Ren, Z. Sha, Z. Fu, Z. Xu, Z. Xie, Z. Zhang, Z. Hao, Z. Ma, Z. Yan, Z. Wu, Z. Gu, Z. Zhu, Z. Liu, Z. Li, Z. Xie, Z. Song, Z. Pan, Z. Huang, Z. Xu, Z. Zhang, and Z. Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

N. Heess, D. TB, S. Sriram, J. Lemmon, J. Merel, G. Wayne, Y. Tassa, T. Erez, Z. Wang, S. M. A. Eslami, M. Riedmiller, and D. Silver. Emergence of locomotion behaviours in rich environments. *arXiv:1707.02286 [cs]*, 07 2017. URL https://arxiv.org/abs/1707.02286.

D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv:2103.03874 [cs]*, 11 2021a. URL https://arxiv.org/abs/2103.03874.

D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt. Measuring mathematical problem solving with the math dataset. *NeurIPS*, 2021b.

A. Jaech, A. Kalai, A. Lerer, A. Richardson, A. El-Kishky, A. Low, A. Helyar, A. Madry, A. Beutel, A. Carney, A. Iftimie, A. Karpenko, A. T. Passos, A. Neitz, A. Prokofiev, A. Wei, A. Tam, A. Bennett, A. Kumar, A. Saraiva, A. Vallone, A. Duberstein, A. Kondrich, A. Mishchenko, A. Applebaum, A. Jiang, A. Nair, B. Zoph, B. Ghorbani, B. Rossen, B. Sokolowsky, B. Barak, B. McGrew, B. Minaiev, B. Hao, B. Baker, B. Houghton, B. McKinzie, B. Eastman, C. Lugaresi, C. Bassin, C. Hudson, C. M. Li, d. Bourcy, C. Voss, C. Shen, C. Zhang, C. Koch, C. Orsinger, C. Hesse, C. Fischer, C. Chan, D. Roberts, D. Kappler, D. Levy, D. Selsam, D. Dohan, D. Farhi, D. Mely, D. Robinson, D. Tsipras, D. Li, D. Oprica, E. Freeman, E. Zhang, E. Wong, E. Proehl, E. Cheung, E. Mitchell, E. Wallace, E. Ritter, E. Mays, F. Wang, F. P. Such, F. Raso, F. Leoni, F. Tsimpourlas, F. Song, v. Lohmann,

F. Sulit, G. Salmon, G. Parascandolo, G. Chabot, G. Zhao, G. Brockman, G. Leclerc, H. Salman, H. Bao, H. Sheng, H. Andrin, H. Bagherinezhad, H. Ren, H. Lightman, H. W. Chung, I. Kivlichan, I. O'Connell, I. Osband, I. C. Gilaberte, I. Akkaya, I. Kostrikov, I. Sutskever, I. Kofman, J. Pachocki, J. Lennon, J. Wei, J. Harb, J. Twore, J. Feng, J. Yu, J. Weng, J. Tang, J. Yu, J. Q. Candela, J. Palermo, J. Parish, J. Heidecke, J. Hallman, J. Rizzo, J. Gordon, J. Uesato, J. Ward, J. Huizinga, J. Wang, K. Chen, K. Xiao, K. Singhal, K. Nguyen, K. Cobbe, K. Shi, K. Wood, K. Rimbach, K. Gu-Lemberg, K. Liu, K. Lu, K. Stone, K. Yu, L. Ahmad, L. Yang, L. Liu, L. Maksin, L. Ho, L. Fedus, L. Weng, L. Li, L. McCallum, L. Held, L. Kuhn, L. Kondraciuk, L. Kaiser, L. Metz, M. Boyd, M. Trebacz, M. Joglekar, M. Chen, M. Tintor, M. Meyer, M. Jones, M. Kaufer, M. Schwarzer, M. Shah, M. Yatbaz, M. Y. Guan, M. Xu, M. Yan, M. Glaese, M. Chen, M. Lampe, M. Malek, M. Wang, M. Fradin, M. McClay, M. Pavlov, M. Wang, M. Wang, M. Murati, M. Bavarian, M. Rohaninejad, N. McAleese, N. Chowdhury, N. Chowdhury, N. Ryder, N. Tezak, N. Brown, O. Nachum, O. Boiko, O. Murk, O. Watkins, P. Chao, P. Ashbourne, P. Izmailov, P. Zhokhov, R. Dias, R. Arora, R. Lin, R. G. Lopes, R. Gaon, R. Miyara, R. Leike, R. Hwang, R. Garg, R. Brown, R. James, R. Shu, R. Cheu, R. Greene, S. Jain, S. Altman, S. Toizer, S. Toyer, S. Miserendino, S. Agarwal, S. Hernandez, S. Baker, S. McKinney, S. Yan, S. Zhao, S. Hu, S. Santurkar, S. R. Chaudhuri, S. Zhang, S. Fu, S. Papay, S. Lin, S. Balaji, S. Sanjeev, S. Sidor, T. Broda, A. Clark, T. Wang, T. Gordon, T. Sanders, T. Patwardhan, T. Sottiaux, T. Degry, T. Dimson, T. Zheng, T. Garipov, T. Stasi, T. Bansal, T. Creech, T. Peterson, T. Eloundou, V. Qi, V. Kosaraju, V. Monaco, V. Pong, V. Fomenko, W. Zheng, W. Zhou, W. McCabe, W. Zaremba, Y. Dubois, Y. Lu, Y. Chen, Y. Cha, Y. Bai, Y. He, Y. Zhang, Y. Wang, Z. Shao, and Z. Li. Openai o1 system card, 2024. URL https://arxiv.org/abs/2412.16720.

N. Lambert, J. Morrison, V. Pyatkin, S. Huang, H. Ivison, F. Brahman, M. Lester, A. Liu, N. Dziri, S. Lyu, Y. Gu, S. Malik, V. Graf, J. D. Hwang, J. Yang, R. L. Bras, O. Tafjord, C. Wilhelm, L. Soldaini, N. A. Smith, Y. Wang, P. Dasigi, and H. Hajishirzi. Tulu 3: Pushing frontiers in open language model post-training, 2024. URL https://arxiv.org/abs/2411.15124.

H. Lee, S. Phatale, H. Mansoor, T. Mesnard, J. Ferret, K. Lu, C. Bishop, E. Hall, V. Carbune, A. Rastogi, S. Prakash, and G. Research. Rlaif: Scaling reinforcement learning from human feedback with ai feedback, 2024. URL https://arxiv.org/pdf/2309.00267.

A. Lewkowycz, A. Andreassen, D. Dohan, E. Dyer, H. Michalewski, V. Ramasesh, A. Slone, C. Anil, I. Schlag, T. Gutman-Solo, Y. Wu, B. Neyshabur, G. Gur-Ari, and V. Misra. Solving quantitative reasoning problems with language models. *arXiv:2206.14858 [cs]*, 06 2022. URL https://arxiv.org/abs/2206.14858.

H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever, and K. Cobbe. Let's verify step by step, 05 2023. URL https://arxiv.org/abs/2305.20050.

Z. Liu, C. Chen, W. Li, P. Qi, T. Pang, C. Du, W. S. Lee, and M. Lin. Understanding r1-zero-like training: A critical perspective, 2025. URL https://arxiv.org/abs/2503.20783.

MAA. American mathematics competitions – mathematical association of america, 2024. URL https://maa.org/student-programs/amc/.

MAA. Maa invitational competitions – mathematical association of america, 2025. URL https://maa.org/maa-invitational-competitions/.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback. *arXiv:2203.02155 [cs]*, 03 2022. URL https://arxiv.org/abs/2203.02155.

D. Saxton, E. Grefenstette, F. Hill, and P. Kohli. Analysing mathematical reasoning abilities of neural models. *arXiv preprint arXiv:1904.01557*, 2019.

J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms, 08 2017. URL https://arxiv.org/abs/1707.06347.

Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, L. Y. K, Y. Wu, and D. Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

N. Stiennon, L. Ouyang, J. Wu, D. M. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. Christiano. Learning to summarize from human feedback. *arXiv:2009.01325 [cs]*, 02 2022. URL https://arxiv.org/abs/2009.01325.

J. Uesato, N. Kushman, R. Kumar, F. Song, N. Siegel, L. Wang, A. Creswell, G. Irving, and I. Higgins. Solving math word problems with process- and outcome-based feedback, 11 2022. URL https://arxiv.org/abs/2211.14275.

P. Wang, L. Li, Z. Shao, X. R. X, D. Dai, Y. Li, D. Chen, Y. Wu, and Z. Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations, 2023. URL https://arxiv.org/abs/2312.08935.

S. Wang, L. Yu, C. Gao, C. Zheng, S. Liu, R. Lu, K. Dang, X. Chen, J. Yang, Z. Zhang, Y. Liu, A. Yang, A. Zhao, Y. Yue, S. Song, B. Yu, G. Huang, and J. Lin. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning, 2025. URL https://arxiv.org/abs/2506.01939.

R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 05 1992. doi: 10.1007/bf00992696.

A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, C. Zheng, D. Liu, F. Zhou, F. Huang, F. Hu, H. Ge, H. Wei, H. Lin, J. Tang, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Zhou, J. Lin, K. Dang, K. Bao, K. Yang, L. Yu, L. Deng, M. Li, M. Xue, M. Li, P. Zhang, P. Wang, Q. Zhu, R. Men, R. Gao, S. Liu, S. Luo, T. Li, T. Tang, W. Yin, X. Ren, X. Wang, X. Zhang, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Zhang, Y. Wan, Y. Liu, Z. Wang, Z. Cui, Z. Zhang, Z. Zhou, and Z. Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Q. Yu, Z. Zhang, R. Zhu, Y. Yuan, X. Zuo, Y. Yue, W. Dai, T. Fan, G. Liu, L. Liu, X. Liu, H. Lin, Z. Lin, B. Ma, G. Sheng, Y. Tong, C. Zhang, M. Zhang, W. Zhang, H. Zhu, J. Zhu, J. Chen, J. Chen, C. Wang, H. Yu, Y. Song, X. Wei, H. Zhou, J. Liu, W.-Y. Ma, Y.-Q. Zhang, L. Yan, M. Qiao, Y. Wu, and M. Wang. Dapo: An open-source llm reinforcement learning system at scale, 2025. URL https://arxiv.org/abs/2503.14476.

C. Zheng, S. Liu, M. Li, X.-H. Chen, B. Yu, C. Gao, K. Dang, Y. Liu, R. Men, A. Yang, J. Zhou, and J. Lin. Group sequence policy optimization, 2025. URL https://www.arxiv.org/abs/2507.18071.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes] Sections 3, 4, and 5.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Not Applicable]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Not Applicable]
   (b) Complete proofs of all theoretical results. [Not Applicable]
   (c) Clear explanations of any assumptions. [Not Applicable]
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes] Section 5, Appendix B.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes] Section 5 (Evaluation).
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes] Appendix B.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes] Section 5, Appendix B.
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Appendix

A Limitations and Future Work

While DISPO demonstrates significant improvements over existing methods, several limitations warrant further investigation. First, our experiments focus primarily on mathematical reasoning tasks where binary reward signals are straightforward to obtain. Extending DISPO to domains with more nuanced reward structures, such as code generation or open-ended dialogue, remains unexplored. Second, although we provide empirical guidelines for setting the four clipping parameters ($\epsilon^{+}_{\text{high}}$, $\epsilon^{+}_{\text{low}}$, $\epsilon^{-}_{\text{high}}$, $\epsilon^{-}_{\text{low}}$), determining optimal values still requires some trial and error. Future work could explore adaptive or learned clipping schedules that automatically adjust these parameters based on training dynamics. Additionally, our analysis reveals that excessive exploration in Regimes 1 and 2 can lead to gradual performance degradation in later training stages. While we mitigate these through careful parameter tuning, developing principled methods to detect and prevent such degradations—perhaps through entropy regularization or dynamic exploration schedules—would be valuable. Finally, due to computational constraints, we limited our experiments to models up to 30B parameters. Future work could investigate DISPO's scalability to larger models as the field moves toward even bigger reasoning systems.

B Experimental Details

B.1 RLVR Baselines

We evaluate DISPO against representatives from both PPO and REINFORCE families of algorithms. We select DAPO as our PPO-style baseline and employ our own CISPO implementation as the REINFORCE-style baseline. Since the original CISPO's clipping parameters remain unpublished, we configure CISPO with $\epsilon_{\text{low}} = 1$ and $\epsilon_{\text{high}} = 100$, settings that maintain stability across all tested models. Notably, reducing these values (setting $\epsilon_{\text{low}} < 1$ or $\epsilon_{\text{high}} < 100$) triggers training collapses in CISPO. For DAPO, we retain the standard clipping parameters: $\epsilon_{\text{low}} = 0.2$ and $\epsilon_{\text{high}} = 0.28$. We focus on a single PPO-style baseline given that prior work demonstrates CISPO's superiority over various PPO methods, including both DAPO and GRPO [Chen et al., 2025].

B.2 Training

We use the chat template shown in Figure 9 for training. Following the DAPO recipe [Yu et al., 2025], we apply dynamic sampling in all training runs. We use a mini-batch size of 512 and a micro-batch size of 32, corresponding to 16 gradient updates per mini-batch. We set the group size to $G = 16$. We do not include a KL divergence regularization term in the loss calculation. Our max response length is 20,480 tokens. We use the same overlong penalty as in the DAPO recipe in all our experiments. We use AdamW optimizer with learning rate of 1×10^{-6} for all models. In AdamW, we set $(\beta_1, \beta_2) = (0.9, 0.95)$, $\epsilon = 1\text{e}{-15}$, and a weight decay of 0.1, following the CISPO recipe [Chen et al., 2025]. We apply gradient norm clipping at 1.0 and use early truncation with repetition detection, as introduced in the CISPO recipe. All models were trained on H200 GPUs using the Verl codebase with vLLM rollout, requiring approximately 20,000 GPU-hours each. During training, we set the temperature to 1.0 and top-p to 1.0, while for inference we used temperature = 1.0 and top-p = 0.7 across all models. For inference on AIME'24, AIME'25, and AMC'23, we used the system prompt shown in Figure 9, while for MATH-500 and Minerva we used the Qwen3-style system prompt shown in Figure 10 to ensure consistent parsing.

B.3 Evaluation

We evaluate our models using five widely recognized mathematical reasoning benchmarks, namely, **AIME'24** [MAA, 2025], **AIME'25** [MAA, 2025], **AMC'23** [MAA, 2024], **MATH-500** [Hendrycks et al., 2021a], and **Minerva** [Lewkowycz et al., 2022]. These benchmarks assess the model's capacity to solve complex problems across diverse domains and levels of difficulty. Each problem requires the generation of a final answer, which is usually a number, a simplified expression (e.g., $p - q$), or a concise textual response (e.g., *even*):

▷ **AIME'24**: 30 problems from the American Invitational Mathematics Examination in 2024 [MAA, 2025].

```
Solve the following math problem step by step. The
last line of your response should be of the form
Answer: $Answer (without quotes) where $Answer is
the answer to the problem.

{question}

Remember to put your answer on its own line after
\"Answer:\".
```

Figure 9: The DAPO-style chat template we used for training.

```
{question}

Please reason step by step, and put your final
answer within \boxed{}.
```

Figure 10: The Qwen3-style chat template we used for evaluation.

Each problem typically requires multiple steps of intricate reasoning and has an answer that is an integer between 0 and 999.

- ▷ **AIME'25:** A forthcoming set of 30 problems from the AIME 2025 exam [MAA, 2025], covering a similar range of topics in algebra, geometry, number theory, and combinatorics.
- ▷ **AMC'23:** 40 problems from the American Mathematics Competitions in 2023 [MAA, 2024]. This exam serves as a precursor to the AIME and includes problems designed to test creative problem-solving across algebra, geometry, number theory, and probability.
- ▷ **MATH-500:** A curated subset of 500 problems drawn from the MATH dataset [Hendrycks et al., 2021a], selected as Lightman et al. [2023]. These problems cover seven subjects: prealgebra, algebra, number theory, counting and probability, geometry, intermediate algebra, and precalculus.
- ▷ **Minerva:** A benchmark introduced by Lewkowycz et al. [2022], consisting of 272 advanced quantitative reasoning problems drawn from diverse sources such as research-level mathematics and science exams.

To evaluate the correctness of the model's outputs, we follow standard practices in mathematical LLM evaluation. We parse each generated solution using regular expressions to extract the final answer and compare it against the ground truth. For every question in the evaluation set, we generate 16 responses and report the average accuracy (denoted as Avg@16). We track the models' Avg@16 accuracy on AIME'24 throughout training and select the checkpoint with the highest accuracy for final evaluation on all benchmarks. In addition, we compute auxiliary metrics—including token-level average entropy and response length—using the same set of 16 generated completions.

C Additional experimental results

Figures 11 and 12 show all learning curves of Qwen3-30B-A3B-Base and Qwen3-8B-Base, respectively. We see similar characteristics in all curves as in Figure 5, showing the robustness of DISPO across different model sizes and architectures.

Figure 13 shows an example of a token that is repeated in inference time, as mentioned in Regime 3 discussion in Section 5.3.

Figure 14 shows the response length curve of Qwen3-14B-Base. The response length curves of Regime 1 and Regime 2 runs are presented in Figures 15 and 16, respectively.

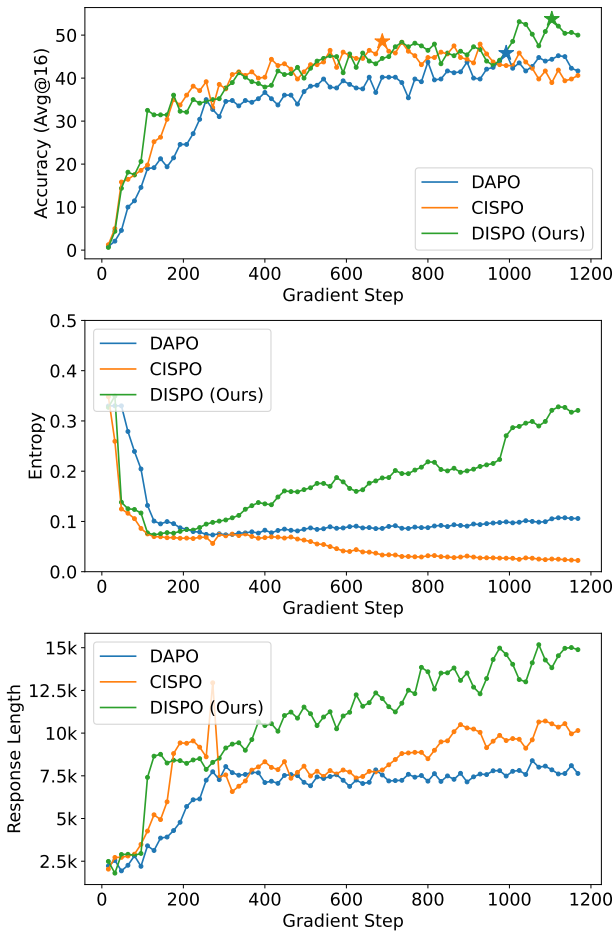

Figure 11: Learning curves of the Qwen3-30B-A3B-Base runs. ⋆ indicates the maximum accuracy.

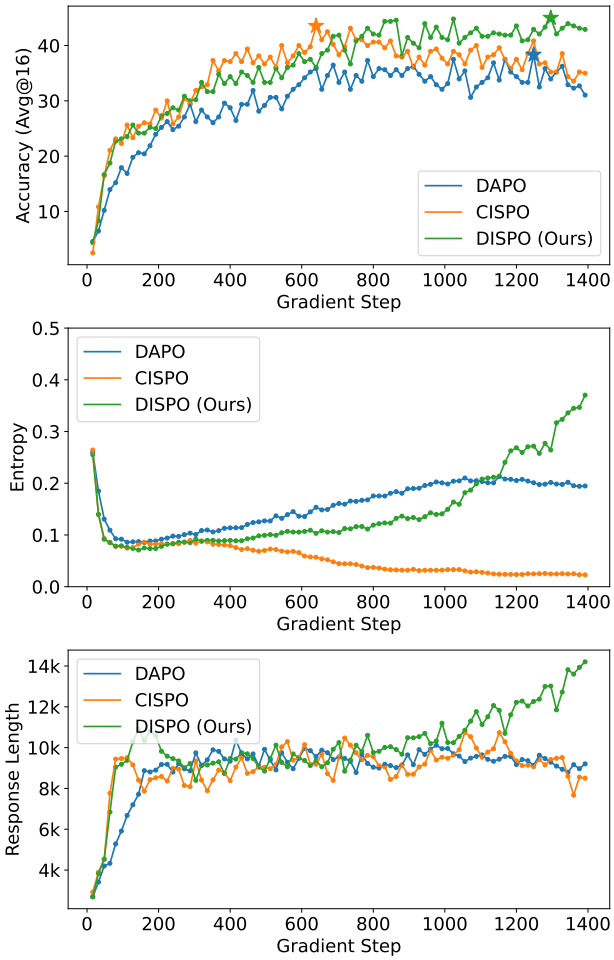

Figure 12: Learning curves of the Qwen3-8B-Base runs. ⋆ indicates the maximum accuracy.

D Additional discussion about Regimes 1 and 2

Figure 17 shows the results when we allow both $r_{i,t} > 1$ and $r_{i,t} < 1$ by setting both $\epsilon^{+}_{\text{low}}$ and $\epsilon^{+}_{\text{high}}$ to be non-zero. This configuration enables the model to simultaneously amplify learning signals (Regime 1) and reduce recovery signals (Regime 2), creating an interplay between exploration and distillation mechanisms and resulting in higher accuracy. The entropy dynamics reveal this competition between Regimes 1 and 2 clearly. The green curve ($\epsilon^{+}_{\text{low}} = 0.28, \epsilon^{+}_{\text{high}} = 1$) in Figure 17 exhibits slower entropy growth compared to the orange curve (Regime 1 only with $\epsilon^{+}_{\text{low}} = 0.28$), demonstrating how the simultaneous activation of Regime 2 counteracts the entropy-increasing effect of Regime 1.

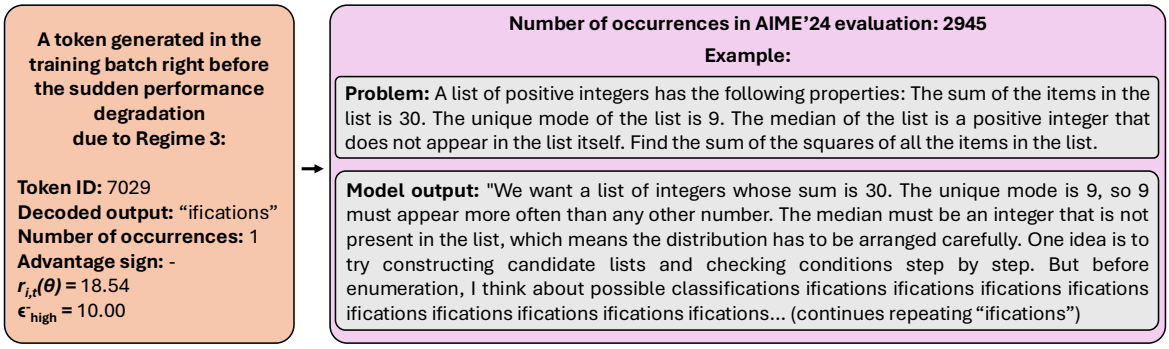

Figure 13: An example of Regime 3 degradation: when $r_{i,t}(\theta) > \epsilon^{-}_{\text{high}}$, Regime 3 update is suppressed, causing the token to become trapped in the high-probability region and repeatedly generated during evaluation.

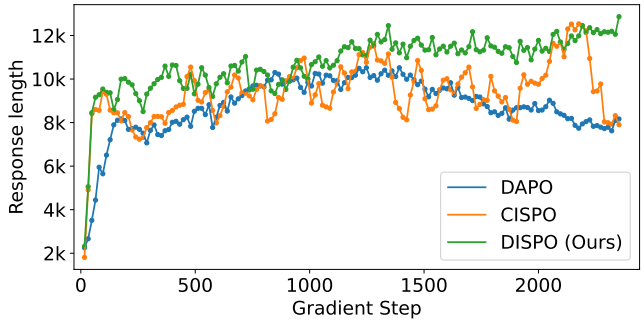

Figure 14: Response length curves of the Qwen3-14B-Base runs.

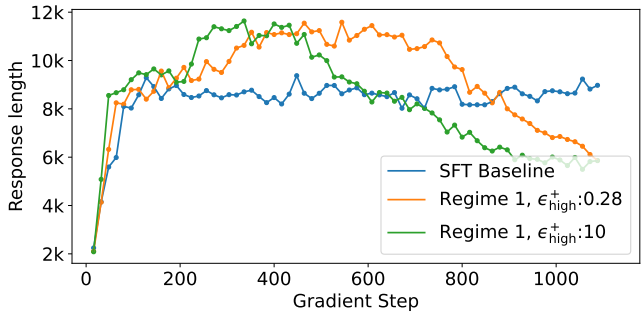

Figure 15: Response length curves of the Regime 1 runs.

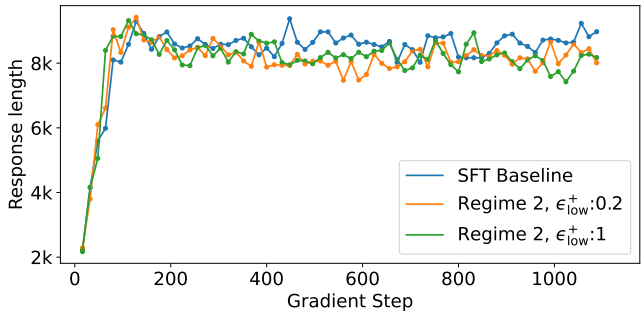

Figure 16: Response length curves of the Regime 2 runs.

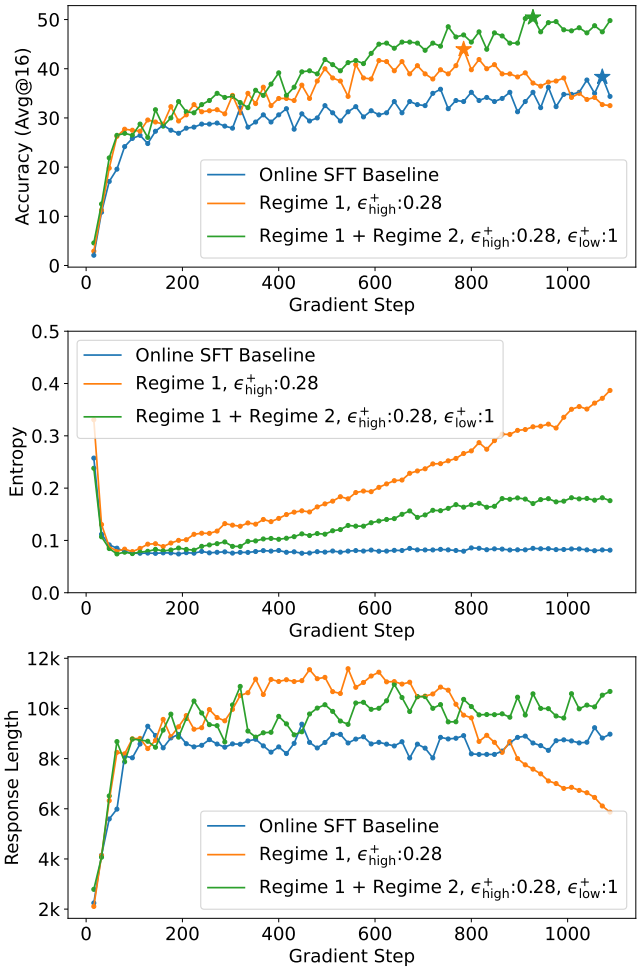

Figure 17: Learning curves of the Regime 1 + Regime 2 run. * in the accuracy panel indicates the maximum value.