RETHINKING ENTROPY INTERVENTIONS IN RLVR: AN ENTROPY CHANGE PERSPECTIVE

Zhezheng Hao[1]* **Hong Wang**[2]* **Haoyang Liu**[2] **Jian Luo**[2]
Jiarui Yu[3] **Hande Dong**[3]† **Qiang Lin**[3] **Can Wang**[1] **Jiawei Chen**[1]†
[1] Zhejiang University [2] University of Science and Technology of China [3] Tencent
{haozhezheng, wcan, sleepyhunt}@zju.edu.cn
{wanghong1700, dgyoung, jianluo}@mail.ustc.edu.cn
{jiaruiyu, handedong, cheaterlin}@tencent.com

ABSTRACT

While Reinforcement Learning with Verifiable Rewards (RLVR) can enhance LLM reasoning, its training process poses a critical risk: entropy collapse. This phenomenon is a rapid loss of policy diversity, stemming from the exploration-exploitation imbalance and leading to a lack of generalization. Recent entropy-intervention methods aim to prevent entropy collapse, yet their underlying mechanisms remain unclear. In this paper, we conduct a quantitative analysis to reveal token-level entropy changes and how existing entropy intervention methods help avoid entropy collapse. Our findings point out a fundamental limitation of existing methods: they attempt to control entropy dynamics indirectly. By only affecting related factors, such as the advantage signal and generation probability, their effectiveness is inherently limited and could potentially fail. To address this limitation, we introduce an entropy-change-aware reweighting scheme, namely **S**tabilizing **T**oken-level **E**ntropy-chang**E** via **R**eweighting (**STEER**), that adaptively stabilizes entropy dynamics through fine-grained token-level adjustments. Our approach mitigates over-exploitation while fostering robust exploration. Extensive experiments demonstrate that **STEER** significantly mitigates entropy collapse, stabilizes entropy dynamics, and achieves stronger downstream performance across various mathematical reasoning benchmarks [1].

1 INTRODUCTION

Reinforcement Learning with Verifiable Rewards (RLVR) has emerged as a powerful paradigm for advancing LLM reasoning (Zhang et al., 2025a; Jaech et al., 2024; Lambert et al., 2024; Guo et al., 2025; Shao et al., 2024; Yang et al., 2025a; Team et al., 2025), with pre-trained LLMs optimized through RLVR demonstrating emergent capabilities such as long-form chain-of-thought and self-reflection (Shao et al., 2024; Zhu et al., 2025b). Despite its success, a key challenge in RLVR remains: the exploration-exploitation trade-off under outcome-based supervision (Yeo et al., 2025; Yue et al., 2025). This is because rewards based solely on the final answer can force models into a state of premature convergence, where models stick to narrow solutions and ignore other correct ones. This issue is particularly damaging to group-based policy-gradient methods (Shao et al., 2024; Ahmadian et al., 2024), as the lack of output diversity makes it difficult to estimate relative advantages, thus providing weak learning signals.

This lack of output diversity is a direct consequence of a poorly managed exploration-exploitation trade-off, a balance often quantified by policy entropy (Wu et al., 2025; Song et al., 2025; Li et al., 2025; Cui et al., 2025b). Low entropy indicates insufficient exploration (a state of over-exploitation), while high entropy indicates sufficient exploration. Therefore, preventing a catastrophic drop in policy entropy, known as entropy collapse, becomes a research focus in RLVR. To avoid entropy col-

*Equal Contribution.

†Corresponding Authors

[1] Our code is available at https://github.com/zz-haooo/STEER.

lapse, existing approaches attempt to indirectly influence entropy dynamics through several mechanisms, each with inherent limitations. One strategy targets **(i)** PPO-style ratio-clipping thresholds, for example, by decoupling them to enhance exploration (Yu et al., 2025); however, this approach can induce asymmetric and uncontrolled effects on entropy change. Another focuses on **(ii)** the relative weighting of positive and negative samples, either by up-weighting rare-but-correct solutions (He et al., 2025a) or skewing weights towards negative samples (Zhu et al., 2025a). While effective at preventing over-exploitation, this method only modulates entropy as a byproduct and lacks fine-grained control. The third approach involves **(iii)** an entropy-induced advantage (Cheng et al., 2025; Tan & Pan, 2025; Wang et al., 2025c;b; Deng et al., 2025). However, this design often has an unintended negative effect: it tends to excessively focus learning on high-entropy tokens, which, instead of stabilizing entropy, amplifies its fluctuations. These observations lead to an important question: **Is there a unified framework that not only explains the root cause of limitations of existing methods, but also guides us to design better solutions?**

We believe the answer is to analyze the problem from the perspective of entropy dynamics. We argue that the overall entropy dynamics during training arises from the accumulation of per-token entropy changes; thus, analyzing entropy change at the token level helps reveal the entropy dynamics. In this paper, we unify the entropy interventions in RLVR through the lens of entropy change: we conduct a quantitative analysis of token-level entropy change, which not only allows us to reveal interesting properties and limitations of existing entropy-intervention methods, but also motivates us to propose a simple yet effective method to control entropy change.

Specifically, we start by conducting a quantitative analysis on token-level entropy change. Based on this analysis, we qualitatively explain how existing methods influence entropy dynamics: (i) PPO-style ratio-clipping thresholds induce asymmetric effects on entropy change; (ii) the relative weighting of positive and negative samples modulates entropy change; and (iii) entropy-induced-advantage approaches magnify entropy fluctuations, which potentially accelerate entropy dropping. Although these methods can influence entropy change, they fall short of controlling entropy change directly. Guided by this insight, we introduce an entropy-change-aware scheme, called **S**tabilizing **T**oken-level **E**ntropy-chang**E** via **R**eweighting (**STEER**), that provides fine-grained, token-level control of policy entropy dynamics to keep per-step entropy change within a moderate band. In this way, our method steers the policy away from over-exploitation and sustains adequate exploration. Empirically, our method achieves superior downstream performance over strong baselines while effectively preventing entropy collapse and strengthening exploration across RLVR benchmarks.

Overall, our contributions can be briefly summarized as follows:

- We propose a quantitative analysis framework for entropy change. Building on this, the effect of entropy interventions can be unified and elucidated through token-level analysis.
- To precisely stabilize entropy change, we propose an adaptive and fine-grained reweighting method that keeps per-step entropy change within a moderate band.
- Experiments on standard RLVR setups demonstrate superior performance, training stability, and precise control of entropy.

2 Preliminaries

2.1 RLVR Algorithms

Given a prompt q sampled from data $\mathcal{D}$, let π_θ denote the policy model parameterized with θ and o denote the response sampled from $\pi_{\text{old}}(\cdot|q)$. PPO (Schulman et al., 2017) optimizes the policy by maximizing the expected advantage and stabilizes the training process through the clipped surrogate. Instead of training an additional value model, GRPO (Shao et al., 2024) samples a group of rollouts $o_{i\,i=1}^{\,G}$ for each prompt q and estimates advantages by relative rewards within the group:

$$A_{i,t} = \frac{R_i - \text{mean}(\{R_i\}_{i=1}^G)}{\text{std}(\{R_i\}_{i=1}^G)}, \tag{1}$$

where R_i equals 1 when the response is correct and -1 when the response is wrong. The advantage $A_{i,t}$ is equivalent for all tokens in the i-th response. Formally, by adapting the token-level policy gradient loss (Yu et al., 2025), GRPO maximizes the following objective.

$$\mathcal{J}(\theta) = \mathbb{E}_{q\sim\mathcal{D},\, \{o_i\}_{i=1}^{G}\sim\pi_{\text{old}}(\cdot|q)} \left[\frac{1}{\sum_{i=1}^{G}|o_i|} \sum_{i=1}^{G}\sum_{t=1}^{|o_i|} \min\left(r_{i,t}A_{i,t}, \text{clip}\left(r_{i,t}, 1+\varepsilon, 1-\varepsilon\right)A_{i,t}\right) \right], \tag{2}$$

where $r_{i,t} = \frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\text{old}}(o_{i,t}|q,o_{i,<t})}$ denotes the importance sampling ratio. The KL divergence term in the original GRPO (Shao et al., 2024) is excluded in this work to enable broader exploration.

2.2 Policy entropy of LLMs

Shannon entropy quantifies the uncertainty of a policy model's action selection given a state (Haarnoja et al., 2018). The token entropy on token $o_{i,t}$ is defined as the entropy of the conditional distribution $\pi_\theta(\cdot|q, o_{i,<t})$:

$$\mathcal{H}_{i,t} = -\mathbb{E}_{o_{i,t}\sim\pi_\theta(\cdot|q,o_{i,<t})}\left[\log \pi_\theta(o_{i,t}|q, o_{i,<t})\right]. \tag{3}$$

Policy entropy measures a policy model's overall uncertainty on a dataset by averaging token entropy over all tokens. For policy model π_θ and the dataset $\mathcal{D}$, the policy entropy is defined as:

$$\mathcal{H}(\pi_\theta, \mathcal{D}) = \mathbb{E}_{q\sim\mathcal{D},\, \{o_i\}_{i=1}^{G}\sim\pi_{\text{old}}(\cdot|q)} \frac{1}{\sum_{i=1}^{G}|o_i|} \sum_{i=1}^{G}\sum_{t=1}^{|o_i|} \mathcal{H}_{i,t}. \tag{4}$$

Although Eq. (4) is dataset-dependent, it reflects the diversity of model's responses within a specific domain (e.g., mathematical reasoning). Thus, we use $\mathcal{D} = \mathcal{D}_{\text{train}}$ when computing policy entropy.

3 Entropy-intervention Mechanism: An Entropy Change Perspective

Policy entropy serves as an indicator of a model's output diversity. The overall entropy change reflects the exploration–exploitation trade-off during training. The overall changes in policy entropy accumulates from micro-level entropy changes, with a single update's effect on a single token's conditional entropy constituting the atomic unit. In this section, we begin from this micro-level perspective, deriving a quantitative analysis to identify the direct factors that govern token-level entropy change. We then leverage this analysis to examine the impact of existing training parameters on the overall entropy dynamics.

3.1 Quantitative Analysis on Token-level Entropy Change

We start by analyzing the factors that govern a single token's entropy change. The policy gradient of GRPO (in Eq. (2)) is expressed as follows:

$$\nabla_\theta J(\theta) = \mathbb{E}_{q\sim\mathcal{D},\, \{o_i\}\sim\pi_{\text{old}}(\cdot|q)} \left[\frac{1}{\sum_{i=1}^{G}|o_i|} \sum_{i=1}^{G}\sum_{t=1}^{|o_i|} \mathbb{I}_{\text{clip}}\, r_{i,t}A_{i,t}\nabla_\theta \log \pi_\theta(o_{i,t} \mid q, o_{i,<t}) \right], \tag{5}$$

where

$$\mathbb{I}_{\text{clip}} = \begin{cases} 0, & A_{i,t} > 0 \ \text{ and } \ r_{i,t} > 1+\varepsilon_{\text{high}}, \\ 0, & A_{i,t} < 0 \ \text{ and } \ r_{i,t} < 1-\varepsilon_{\text{low}}, \\ 1, & \text{otherwise}. \end{cases} \tag{6}$$

During the RLVR training process, token-level logit distributions are influenced by multiple factors, so it is impractical to estimate the induced entropy change in entropy directly. To capture the essence of distribution shifts, we follow the assumption from (Liu, 2025):

Assumption 1 (Parameter-independent softmax). *For any context (state) $s_{i,t} = (q, o_{i,<t})$, each token (action) a in vocabulary $\mathcal{V}$ is associated with an independent logit parameter $z_{s,a}(\theta)$. And the next-token distribution follows $\pi_\theta^k(\cdot \mid s) = \text{softmax}(z_{s,\cdot}^k)$.*

Assumption 1 states that a gradient step on the sampled token does not substantially affect the logits of the other tokens in the vocabulary. Given this assumption, we obtain the following theorem (see proof in Appendix C).

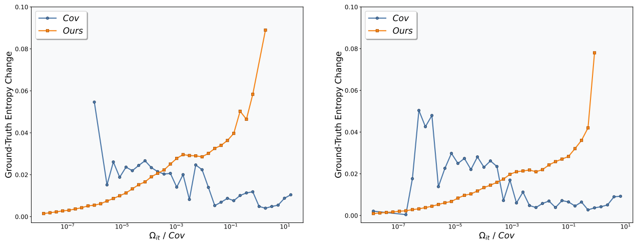
Theorem 1. *(First–order entropy change) Let the policy model π_θ follow Assumption 1. The change of conditional entropy between two update steps is defined as $\Delta\mathcal{H}_{it} \triangleq \mathcal{H}(\pi_\theta^{k+1} \mid s_{i,t}) - \mathcal{H}(\pi_\theta^k \mid s_{i,t})$. Then the first-order estimation of $\Delta\mathcal{H}_{it}$ in Eq. 2 is*

$$\Omega_{i,t} = -\eta\, \mathbb{E}_{a\sim\pi_\theta^k(\cdot|s_{i,t})}\, w_{i,t}(1-\pi_\theta^k(a|s_{i,t}))^2\,(\log \pi_\theta^k(a|s_{i,t}) + \mathcal{H}(\pi_\theta^k \mid s_{i,t})), \tag{7}$$

where η is the learning rate, $w_{i,t} = \mathbb{I}_{clip}\, r_{i,t}\, A_{i,t}$ is per-token weight.

Theorem 1 above implies that, under Assumption 1, the entropy change of a single token $\Delta\mathcal{H}_{it}$ can be reflected by $\Omega_{i,t}$. Obviously, $\Omega_{i,t}$ are jointly determined by learning rate η, per-token gradient weight $w_{i,t}$, generation probability $\pi_\theta^k(a|s_{i,t})$ and current conditional entropy $\mathcal{H}(\pi_\theta^k \mid s_{i,t})$.

In contrast to our Assumption 1, prior work often relies on more restrictive assumptions to derive entropy change. For instance, (Cui et al., 2025b) (denoted as *Cov*) assumes a uniform entropy distribution across different queries within the same batch. However, this assumption is rarely attainable in practice and can lead to estimations that misrepresent the ground-truth entropy dynamics.



(a) Qwen2.5-7B

(b) Qwen2.5-Math-7B

Figure 1: Entropy change estimation in the first 10 training steps on Qwen2.5-7B and Qwen2.5-Math-7B. The curve compares estimated vs. ground-truth entropy change (left axis) and histograms show token counts per bin (right axis).

To validate our approach, we compare our entropy change estimator, $\Omega_{i,t}$, with that of *Cov* during a standard GRPO training process. As visualized in Figure 1, our proposed $\Omega_{i,t}$ closely tracks the ground-truth entropy change, showing a positive correlation. While the estimation *Cov* shows only a weak correlation.

To quantify this gap, we compute the Mean Squared Error (*MSE*), Pearson Correlation Coefficient (*PCC*), and Spearman's Rank Correlation Coefficient (*SRCC*) between each estimation and the ground-truth token-level entropy change, as shown in Figure 2. Across all three metrics, $\Omega_{i,t}$ from Theorem 1 delivers orders-of-magnitude lower MSE and substantially higher *PCC* and *SRCC* than *Cov*. Furthermore, the SRCC between $\Omega_{i,t}$ and the ground-truth token entropy change exceeds 60% across all models, demonstrating a strong rank correlation. A more comprehensive comparison is provided in Appendix D.1. These results strongly validate the effectiveness of our estimator derived in Theorem 1 and the soundness of Assumption 1.

Model	*Method*	*MSE* ↓	*PCC* ↑	*SRCC* ↑
Qwen-Math-1.5B	*Cov*	5.37	-6e-5	+0.04
	Ours	5e-4	+0.42	+0.65
Qwen-7B	*Cov*	0.53	+0.05	+0.08
	Ours	8e-4	+0.39	+0.72
Qwen-Math-7B	*Cov*	0.29	+0.03	+0.06
	Ours	4e-4	+0.42	+0.61

Figure 2: MSE, PCC and SRCC comparison.

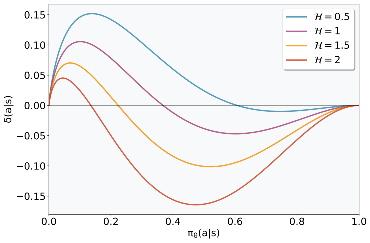


Figure 3: Token-level entropy change indicator $\delta(a|s)$.

3.2 On Analysis of Phenomena in Entropy Dynamics

3.2.1 Entropy Dynamics under Advantage and Probability

To dissect the factors governing token-level entropy change, we first need to decompose the first-order estimation $\Omega_{i,t}$ from Theorem 1. To this end, we define a **token-level entropy change indicator** $\delta(a|s)$ as:

$$\delta(a|s) = -\pi_\theta(a|s)(1-\pi_\theta(a|s))^2(\log(\pi_\theta(a|s)) + \mathcal{H}(\cdot|s)) \tag{8}$$

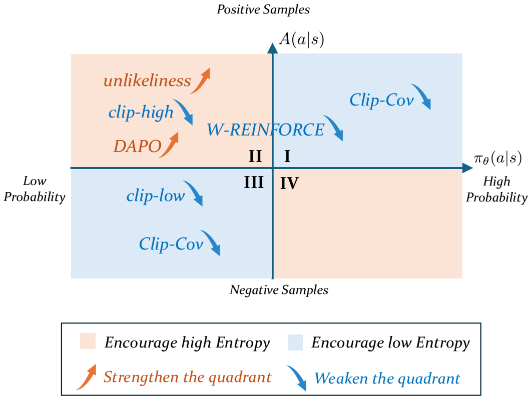


Figure 4: Entropy change with advantage and probability.

Method	$\pi_\theta(a\|s)$	$A(a\|s)$	$\mathcal{H}(\cdot\|s)$
DAPO	✓	✓	✗
Unlikeliness	✓	✓	✗
W-REINFORCE	✗	✓	✗
Entropy Adv.	✗	✓	✓
KL Reg.	✓	✗	✗
Entropy Reg.	✗	✗	✓
Edge-GRPO	✗	✗	✓
Forking Tokens	✗	✗	✓
Clip-Cov	✓	✓	✗
STEER	✓	✓	✓

Figure 5: Key Considerations in Current Approaches.

This allows us to express the entropy change from Theorem 1 as

$$\Omega_{i,t} = \eta\,\mathbb{E}_{a\sim\pi_{\text{old}}(\cdot|s_{i,t})}\Big[\frac{\mathbb{I}_{\text{clip}}\,A(a|s_{i,t})}{\pi_{\text{old}}(a|s_{i,t})}\cdot\delta(a|s_{i,t})\Big]. \tag{9}$$

The key insight is that $\delta(a|s)$ represents the *intrinsic directional tendency* of the entropy change, since it only depends on the token's generation probability $\pi_\theta(a|s)$ and the current conditional entropy $\mathcal{H}(\cdot|s)$. Figure 3 visualizes $\delta(a|s)$ as a function of these two variables.

Based on this decomposition, we can now analyze the entropy dynamics by examining how token-level entropy changes with different signs of the advantage $A(a|s)$ and the indicator $\delta(a|s)$. To illustrate, we create a two-dimensional space, shown in Figure 4, which can be divided into four distinct quadrants:

Quadrant I: Exploitation (entropy decrease). For high-probability correct tokens ($A > 0, \delta < 0$), rewarding an already-mastered behavior concentrates probability mass, thus *decreasing* entropy.

Quadrant II: Exploration (entropy increase). For low-probability correct tokens ($A > 0, \delta > 0$), rewarding a rare-but-correct behavior diversifies the policy, thereby *increasing* entropy.

Quadrant III: Suppression (entropy decrease). For low-probability incorrect tokens ($A < 0, \delta > 0$), penalizing an unlikely error pushes its probability further toward zero, which also *decreases* entropy.

Quadrant IV: Error-Correction (entropy increase). For high-probability incorrect tokens ($A < 0, \delta < 0$), penalizing an overconfident error flattens the distribution to encourage seeking alternatives, substantially *increasing* entropy.

To validate these theoretical findings, we conduct an experiment to provide empirical support. Specifically, based on the above analyses, we can learn that entropy increases in two of these quadrants: (Quadrant II) when updating on low-probability tokens with positive advantages, and (Quadrant IV) when updating on high-probability tokens with negative advantages. To test this, we selectively apply double-weighting (to strengthen) or masking (to weaken) to 10% of tokens falling into each quadrant and track the resulting entropy. As shown in Figure 6, all four interventions successfully increase policy entropy compared to the standard GRPO baseline, confirming our model's validity. Further empirical studies and experimental details are provided in the Appendix D.2.

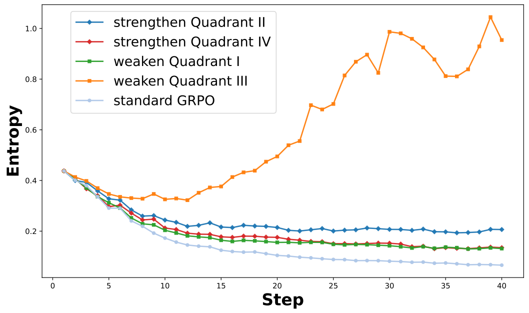
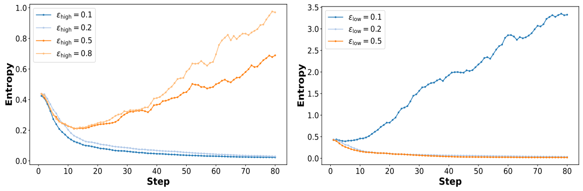


Figure 6: Four schemes to uplift entropy based on advantage and probability.

In a standard RLVR process, these four dynamics co-exist, acting as competing forces that shape the policy. Policy entropy evolves from the superposition of these updates. Consequently, **entropy collapse** can be understood as a state where the exploitation-driven, entropy-decreasing updates (Quadrants I and III) consistently overwhelm the exploration-driven, entropy-increasing updates (Quadrants II and IV). This framework not only explains the phenomenon but also provides a foundation for analyzing the effects of other interventions, such as positive/negative sample rebalancing and ratio clipping.



(a) clip-high

(b) clip-low

Figure 7: Entropy dynamics with ratio clipping.

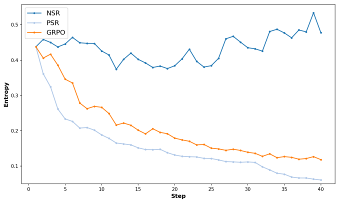


Figure 8: PSR-NSR.

3.2.2 Explaining the Asymmetric Impact of Ratio Clipping

Ratio clipping is a core component of PPO-style algorithms, designed to prevent destructive policy updates by constraining the importance sampling ratio r_t. This mechanism can be interpreted within our framework as a gate that primarily suppresses updates for tokens with large ratios—namely, low-probability tokens. As our analysis in Section 3.2.1 shows, these tokens correspond to the entropy-increasing Quadrant II (exploration) and the entropy-decreasing Quadrant III (suppression).

This insight allows us to form a clear hypothesis about how adjusting the clipping thresholds, $\varepsilon_{\text{high}}$ and ε_{low}, will asymmetrically affect policy entropy:

Adjusting $\varepsilon_{\text{high}}$: This threshold gates updates on positive-reward tokens. Increasing $\varepsilon_{\text{high}}$ (as in DAPO, (Yu et al., 2025)) relaxes the constraint on Quadrant II updates. This should unleash more of the natural, entropy-increasing effect of exploration. We therefore predict that **a higher $\varepsilon_{\text{high}}$ will increase policy entropy**.

Adjusting ε_{low}: This threshold gates updates on negative-reward tokens. Increasing ε_{low} relaxes the constraint on Quadrant III updates. This should amplify the natural, entropy-decreasing effect of suppression. We therefore predict that **a higher ε_{low} will decrease policy entropy**.

To verify our predictions, we conducted two experiments. First, we confirmed that clipping is indeed concentrated on low-probability tokens, as shown by the trigger counts in Figure 9. Second, we independently varied $\varepsilon_{\text{high}}$ and ε_{low} and tracked the resulting entropy dynamics. The results, presented in Figures 7a and 7b, perfectly align with our predictions: entropy rises with a higher $\varepsilon_{\text{high}}$ and falls with a higher ε_{low}.

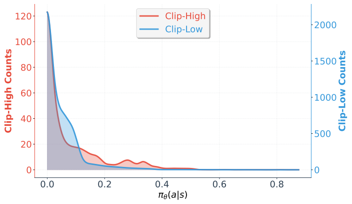


Figure 9: The average clip counts over the first 10 steps.

This analysis demonstrates that our framework provides a principled explanation for the asymmetric and often counter-intuitive effects of ratio clipping on policy entropy. As a consequence, this heuristic entropy intervention method relying on global parameters may avert collapse but instead induces an entropy explosion, which leads to excessive model perplexity.

3.2.3 Explaining the Impact of Positive and Negative Sample Weighting

A notable phenomenon, observed by Zhu et al. (2025a) and confirmed in our experiments (Figure 8), is that training exclusively on negative samples (Negative Sample Reweighting, or NSR) sustains high policy entropy, whereas training only on positive samples (Positive Sample Reweighting, or PSR) leads to a rapid entropy collapse. Our four-quadrant framework provides a clear explanation for this behavior.

The key insight is that training data is naturally dominated by high-probability tokens. While our analysis shows that the *magnitude* of entropy change, $|\delta(a|s)|$, is similar for both high- and low-probability tokens (Figure 3), the sheer volume of high-probability tokens means they dictate the overall entropy trend.

In PSR (Positive Sample Reweighting), the training signal is dominated by high-probability correct tokens, which fall into Quadrant I (Exploitation). This leads to a relentless decrease in entropy. Crucially, PSR removes all negative samples, thereby eliminating the powerful, entropy-increasing force of Quadrant IV (Error-Correction). Without this countervailing force, the policy quickly converges to a narrow solution set, causing entropy to collapse.

In NSR (Negative Sample Reweighting), the training signal is dominated by high-probability incorrect tokens, which fall into Quadrant IV (Error-Correction). This provides a strong and continuous entropy-increasing signal. By removing all positive samples, NSR also eliminates the primary source of entropy decrease from Quadrant I (Exploitation). The result is a policy that constantly seeks to correct its errors, thereby maintaining high diversity and high entropy.

This framework also clarifies the mechanism behind other related methods. For instance, the strategy of up-weighting rare-but-correct tokens, as proposed by He et al. (2025a) and Deng et al. (2025), can be understood as a targeted intervention to boost the entropy-increasing effect of Quadrant II (Exploration). By amplifying this specific signal, these methods aim to counteract the dominant entropy-decreasing pressure from Quadrant I and thus mitigate entropy collapse. Overall, Figure 4 summarizes the entropy effects of some methods on the four quadrants.

3.2.4 The Perils of Targeting High-Entropy Tokens

While advantage and token probability determine the *direction* of an entropy update, the current conditional entropy, $\mathcal{H}(\cdot|s)$, governs its *magnitude*. Our analysis of the entropy change indicator $\delta(a|s)$ reveals a critical dynamic: the magnitude of potential entropy change, $|\delta(a|s)|$, increases significantly as $\mathcal{H}(\cdot|s)$ grows, particularly for high-probability tokens (Figure 3, right half). This implies that tokens in states of high uncertainty are inherently volatile and prone to large swings in entropy. This relationship is empirically confirmed in Figure 10, which shows a strong correlation between a token's current entropy and the magnitude of its subsequent entropy change.

This volatility has led some methods, such as Entro. Adv. (Cheng et al., 2025) and GTPO (Tan & Pan, 2025), to propose interventions that explicitly up-weight high-entropy tokens. The intuition is that focusing on these uncertain states will promote exploration and thus increase overall policy entropy. Nevertheless, our analysis reveals this strategy to be counterproductive and potentially harmful. High-entropy tokens are not a reliable source of entropy *increase*; they are a source of entropy *fluctuation*. By amplifying updates on these tokens, these methods create an unfavorable positive feedback: when policy entropy happens to decrease, the amplified updates on the now high-entropy tokens can cause it to decrease even faster; This creates a system that is highly sensitive to its own fluctuations: instead of stabilizing entropy, it amplifies its inherent fluctuations.

We demonstrate this destabilizing effect in Figure 11. Compared to the standard GRPO baseline, entropy-induced advantage methods exhibit much larger fluctuations. Critically, when the policy enters a phase of decline, these methods can **accelerate entropy collapse**, leading to a faster and more severe drop in diversity. This finding highlights a key flaw in targeting high-entropy tokens: rather than preventing collapse, such interventions can inadvertently aggravate it. This further confirms why (Zhang et al., 2025c) is effective in mitigating entropy collapse.

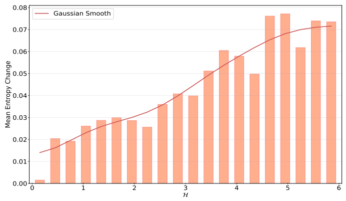
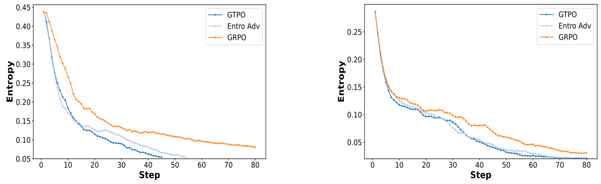


Figure 10: Empirical correlation between current entropy and entropy change.



(a) Math-7B on DAPO-17k

(b) Math-7B on Math

Figure 11: Entropy dynamics with advantage shaping.

4 Stabilizing Token-level Entropy-change via Reweighting

Building on the above analysis, we find that all three factors materially shape entropy change, whereas existing approaches target only a subset, which limits their effectiveness, as shown in Table 5. Since excessive entropy change can cause the policy entropy to rapidly increase or decrease, potentially leading to model training failure, we aim to keep the stepwise entropy change within a moderate range. To control entropy change precisely, we introduce an adaptive and fine-grained token-reweighting scheme that keeps the stepwise entropy change within a moderate band. Since $\Omega_{i,t}$ (Figure 2) shows a strong correlation with the ground-truth entropy change, a simple approach is to design a token-level weight negatively correlated with $\Omega_{i,t}$ to suppress updates of tokens with excessively large entropy changes. Specifically, we apply an exponential-decay mapping to the token weights:

$$\lambda_{i,t} = e^{-k\cdot|\Omega_{i,t}|}, \text{ where } k = \frac{-\ln \lambda_{\min}}{\max\{|\Omega_{i,t}| \mid o_i \in \mathcal{B}\}}, \tag{10}$$

so that the token with the largest entropy change in each mini-batch $\mathcal{B}$ attains the minimum weight. $\lambda_{\min}$ is the only hyperparameter introduced and constrains token weights within $[\lambda_{\min}, 1]$. When $\lambda_{\min}$ equals 1, STEER degenerates into standard GRPO. As shown in Figure 5, compared to existing methods that impact entropy change based on a subset of considerations, STEER takes a more comprehensive and precise approach by considering all relevant factors influencing entropy change. Besides, by controlling entropy change, STEER mitigates entropy collapse without driving training into its symmetric counterpart—entropy explosion. It is noteworthy that this reweighting scheme does not hinder the model's learning, as the reweighting is dominated by a few tokens with very large $|\Omega_{i,t}|$ within the batch, while the majority of tokens keep weights near 1.

5 Experiments

5.1 RLVR training Setups

Training: We conduct experiments on three different models, including Qwen2.5-Math-7B, Qwen2.5-Math-1.5B and Qwen2.5-14B. We adapt our training codebase from verl (Sheng et al., 2025) and follow the training recipe of standard GRPO. Our training data is DAPO-Math-17k (Yu et al., 2025), containing only math problems with integer ground-truth answers. Both the KL-divergence and entropy loss terms are removed in our experiments. Generation batch size is set to 512, and update batch size is set to 32. The number of rollouts is set to 8. Training is performed with top-p value of 1.0 and temperature= 1.0. Training details of our method and baselines are in Appendix E.

Evaluation: We evaluate our models and baselines on six widely used mathematical reasoning benchmarks: AIME24, AIME25, AMC23 (Li et al., 2024), MATH-500 (Hendrycks et al., 2021), Minerva Math (Lewkowycz et al., 2022), and OlympiadBench (He et al., 2024), detailed in Appendix E. Validation is performed with a top-p value of 0.7 and temperature= 1.0 across all models and test sets. We use Math-Verify and Qwen-Verify for both validation during training and final evaluation. All evaluations are *zero-shot* with no additional prompts.

Baselines: For a thorough comparison, we compare our method against 10 baselines, including standard GRPO (Shao et al., 2024), SimpleRL-Zoo (Zeng et al., 2025), Eurus-PRIME (Cui et al., 2025a), OPO (Hao et al., 2025), GRPO with clip-high (Yu et al., 2025), GRPO with entropy loss (Schulman et al., 2017), GRPO with Fork Tokens (Wang et al., 2025c), W-REINFORCE (Zhu et al., 2025a), Entro. Adv. (Cheng et al., 2025), Clip-Cov and KL-Cov (Cui et al., 2025b). For all baselines, the default training hyperparameters are consistent with STEER, while the newly introduced hyperparameters follow the original implementations, respectively.

5.2 Results and Analysis

Main Results: As shown in Table 1, STEER outperforms classical RLVR baselines as well as existing entropy intervention baselines across all datasets. STEER improves average performance by 2.7 points over the second runner-up (OPO) and by 3.4 points over the third runner-up (Clip-Cov) across all baselines. The performance experiments on Qwen2.5-Math-1.5B and Qwen2.5-14B

Table 1: Benchmark results of different methods. We report avg@32 for AIME24, AIME25, and AMC23 and avg@1 for others. All results are presented as percentages.

Method	**AIME24**	**AIME25**	**AMC23**	**MATH500**	**Minerva**	**Olympiad**	**Avg.**
Qwen2.5-Math-7B	13.8	5.3	44.6	39.6	9.9	13.8	21.2
Classical RLVR Baselines							
GRPO	28.0	14.3	66.2	78.6	37.3	40.9	44.2
SimpleRL-Zoo	25.2	13.4	70.6	78.6	37.8	38.4	44.0
Eurus-PRIME	20.9	13.0	65.2	79.8	37.4	40.6	42.8
OPO	32.2	13.4	71.5	82.2	38.2	41.0	46.4
Entropy Intervention Baselines							
GRPO w/ clip-high	31.7	12.8	66.8	79.0	38.6	39.3	44.7
GRPO w/ Entro. Loss	29.1	14.0	67.6	80.0	38.2	37.9	44.5
GRPO w/ Fork Tokens	31.9	14.3	65.5	79.2	37.1	40.9	44.8
W-REINFORCE	31.9	14.3	65.5	79.2	37.1	40.9	44.8
Entro. Adv.	27.5	13.5	70.2	79.6	36.8	42.8	45.1
Clip–Cov	32.5	12.9	68.4	78.0	40.8	41.3	45.7
KL–Cov	32.8	14.1	64.2	78.8	37.1	39.4	44.4
Our Method							
STEER	**36.9**	**16.2**	**72.2**	**82.4**	**41.7**	**43.3**	**49.1**

shown in Figure 5 are compared with the top three competitors in Table 1 (i.e., OPO, Clip-Cov, and Entro. Adv.). STEER also consistently achieves the highest average performance on both Qwen2.5-Math-1.5B (38.1) and Qwen2.5-14B (45.1), demonstrating its superior capabilities in improving model reasoning. Figure 12 shows the test curves during training, where STEER outperforms the baselines. Figure 13 presents the test curves for different hyperparameters, demonstrating both stability and superiority.

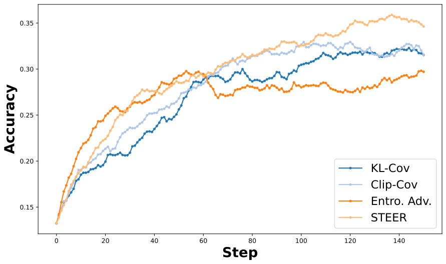


Figure 12: Test set accuracy dynamics comparison with benchmarks.

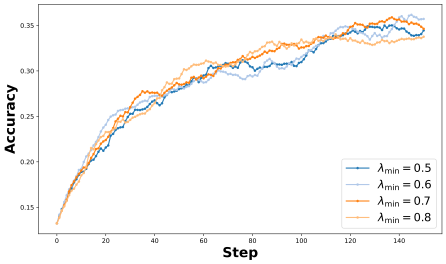


Figure 13: Test set accuracy dynamics comparison under different $\lambda_{\min}$.

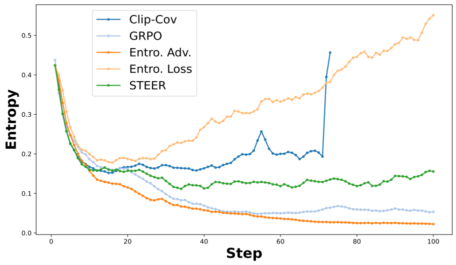


Figure 14: Entropy dynamics in extreme scenarios with benchmarks.

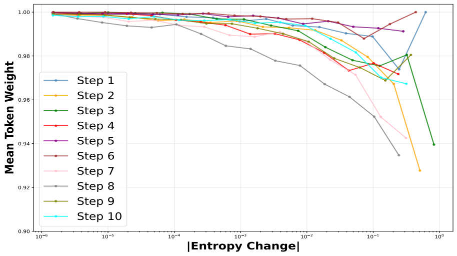


Figure 15: Relationship between mean token weight and entropy change across steps.

Entropy Control: The strength of our method is not only reflected in its performance but also in its ability to regulate entropy across a wide range. We consider an extreme training setup with $\varepsilon_{\text{high}} = 5$ and $\varepsilon_{\text{low}} = 0.99$, where almost no ratio clipping is applied. In such scenarios, RL training is vulnerable due to unstable gradient updates under extreme clipping ratios. The results are shown in the Figure 14. Most methods fail to maintain stable entropy: GRPO and Entro. Adv. tend toward entropy collapse; adding an Entropy Loss drives entropy up rapidly, leading to excessive uncertainty; and Clip-Cov cannot reliably control entropy. By contrast, STEER stabilizes after an initial decline and maintains steady entropy subsequently. The test set results for extreme scenarios are provided in the Appendix D.3. Besides, Figure 15 depicts the trend of average token weight as a function of the absolute token entropy change in the first 10 steps. When entropy changes are small, most weights remain near 1; only tokens with large entropy changes receive substantially reduced weights, indicating that STEER stabilizes training without impeding learning.

Ablation Study: Besides the exponential mapping in Eq. (10), we consider the following linear mapping and binary mapping for ablation:

$$\text{linear:} \quad \lambda_{i,t} = \lambda_{\text{max}} - \frac{\lambda_{\text{max}} - \lambda_{\text{min}}}{\Omega_{\text{max}} - \Omega_{\text{min}}}(\Omega_{i,t} - \Omega_{\text{min}}), \quad \text{binary:} \quad \lambda_{i,t} = \begin{cases} \lambda_{\text{min}}, & \Omega_{i,t} > Q_{\xi}(\Omega), \\ 1, & \text{otherwise.} \end{cases}$$

We set $\lambda_{\text{min}} = 0.7$ throughout and $\lambda_{\text{max}} = 1.2$ for the linear mapping. For the binary mapping, the quantile threshold Q_ξ is set to $Q_{0.8}$—i.e., the top 20% of tokens by $\Omega_{i,t}$ are assigned weight λ_{min} and the remaining 80% of tokens keep weight 1. The three mapping schematics are illustrated in Figure 16, and their performance on Qwen2.5-Math-7B is reported in Table 2. It can be seen that the binary mapping degrades performance, whereas the linear mapping does not materially harm performance. This highlights the necessity of continuous token-level reweighting, as truncation cannot precisely control entropy change.

Table 2: Ablation study on different weight mapping modes.

Mapping	**AIME24**	**AIME25**	**AMC23**	**MATH500**	**Minerva**	**Olympiad**	**Avg.**
exponential	36.9	16.2	72.2	82.4	41.7	43.3	48.8
linear	35.7	15.5	73.6	81.0	39.8	41.9	47.9
binary	32.3	14.4	71.5	82.2	38.3	41.0	46.6

We also assess the sensitivity of the experimental results to hyperparameters λ_{min} in Eq. (10). An excessively small λ_{min} may hinder the model's learning and lead to unstable training, while an excessively large λ_{min} reduces the model's ability to control entropy. As shown in Figure 17, our method performs consistently well when $\lambda_{\text{min}} \in [0.6, 0.8]$.

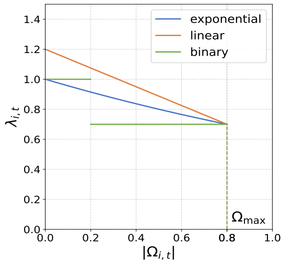


Figure 16: Weight Mapping.

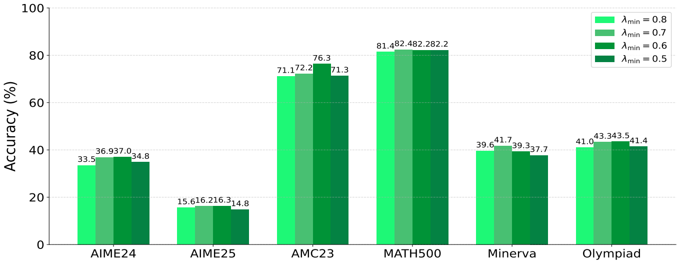


Figure 17: Hyperparameter Sensitivity on λ_{min}.

6 Conclusion

In this paper, we rethink the entropy interventions through the lens of entropy change. By proposing a quantitative analysis framework for entropy change, the entropy effect of current intervention methods can be unified and elucidated through token-level analysis. Motivated by stabilizing entropy change, we propose STEER, an adaptive, fine-grained reweighting scheme that precisely keeps per-step entropy changes within a moderate band by suppressing potentially disruptive updates. Extensive experiments on mathematical reasoning benchmarks demonstrate that STEER achieves superior performance with enhanced training stability. Our work provides both a new lens for analyzing RL dynamics and a practical solution for developing robust and effective LLM training algorithms.

Ethics statement

We have manually reevaluated the dataset we created to ensure it is free of any potential for discrimination, human rights violations, bias, exploitation, and any other ethical concerns.

Reproducibility statement

To ensure the reproducibility of our findings, all source code and datasets used in our experiments are included in the supplementary material. The provided materials are sufficient to replicate the main results presented in this paper.

References

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms. *arXiv preprint arXiv:2402.14740*, 2024.

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

Xiangxiang Chu, Hailang Huang, Xiao Zhang, Fei Wei, and Yong Wang. Gpg: A simple and strong reinforcement learning baseline for model reasoning. *arXiv preprint arXiv:2504.02546*, 2025.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, et al. Process reinforcement through implicit rewards. *arXiv preprint arXiv:2502.01456*, 2025a.

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

Jia Deng, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, and Ji-Rong Wen. Decomposing the entropy-performance exchange: The missing keys to unlocking effective reinforcement learning. *arXiv preprint arXiv:2508.02260*, 2025.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pp. 1861–1870. Pmlr, 2018.

Yaru Hao, Li Dong, Xun Wu, Shaohan Huang, Zewen Chi, and Furu Wei. On-policy rl with optimal reward baseline. *arXiv preprint arXiv:2505.23585*, 2025.

Andre He, Daniel Fried, and Sean Welleck. Rewarding the unlikely: Lifting grpo beyond distribution sharpening. *arXiv preprint arXiv:2506.02355*, 2025a.

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

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

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

Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. *arXiv preprint arXiv:2503.24290*, 2025.

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

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.

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

Jia Li, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Huang, Kashif Rasul, Longhui Yu, Albert Q Jiang, Ziju Shen, et al. Numinamath: The largest public dataset in ai4maths with 860k pairs of competition math problems and solutions. *Hugging Face repository*, 13(9):9, 2024.

Qingbin Li, Rongkun Xue, Jie Wang, Ming Zhou, Zhi Li, Xiaofeng Ji, Yongqi Wang, Miao Liu, Zheming Yang, Minghui Qiu, et al. Cure: Critical-token-guided re-concatenation for entropy-collapse prevention. *arXiv preprint arXiv:2508.11016*, 2025.

Jiacai Liu. How does rl policy entropy converge during iteration? `https://zhuanlan.zhihu.com/p/28476703733`, 2025. Zhihu Column.

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

Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Y Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Tianjun Zhang, Li Erran Li, et al. Deepscaler: Surpassing o1-preview with a 1.5 b model by scaling rl. *Notion Blog*, 2025.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pp. 1928–1937. PmLR, 2016.

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

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

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

Yuda Song, Julia Kempe, and Remi Munos. Outcome-based exploration for llm reasoning. *arXiv preprint arXiv:2509.06941*, 2025.

Hongze Tan and Jianfei Pan. Gtpo and grpo-s: Token and sequence-level reward shaping with policy entropy. *arXiv preprint arXiv:2508.04349*, 2025.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025.

Haozhe Wang, Qixin Xu, Che Liu, Junhong Wu, Fangzhen Lin, and Wenhu Chen. Emergent hierarchical reasoning in llms through reinforcement learning. *arXiv preprint arXiv:2509.03646*, 2025a.

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

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, et al. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.01939*, 2025c.

Fang Wu, Weihao Xuan, Ximing Lu, Zaid Harchaoui, and Yejin Choi. The invisible leash: Why rlvr may not escape its origin. *arXiv preprint arXiv:2507.14843*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

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

Edward Yeo, Yuxuan Tong, Morry Niu, Graham Neubig, and Xiang Yue. Demystifying long chain-of-thought reasoning in llms. *arXiv preprint arXiv:2502.03373*, 2025.

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

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837*, 2025.

Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun Ma, and Junxian He. Simplerl-zoo: Investigating and taming zero reinforcement learning for open base models in the wild. *arXiv preprint arXiv:2503.18892*, 2025.

Kaiyan Zhang, Yuxin Zuo, Bingxiang He, Youbang Sun, Runze Liu, Che Jiang, Yuchen Fan, Kai Tian, Guoli Jia, Pengfei Li, et al. A survey of reinforcement learning for large reasoning models. *arXiv preprint arXiv:2509.08827*, 2025a.

Ruipeng Zhang, Ya-Chien Chang, and Sicun Gao. When maximum entropy misleads policy optimization. *arXiv preprint arXiv:2506.05615*, 2025b.

Xingjian Zhang, Siwei Wen, Wenjun Wu, and Lei Huang. Edge-grpo: Entropy-driven grpo with guided error correction for advantage diversity. *arXiv preprint arXiv:2507.21848*, 2025c.

Xinyu Zhu, Mengzhou Xia, Zhepei Wei, Wei-Lin Chen, Danqi Chen, and Yu Meng. The surprising effectiveness of negative reinforcement in llm reasoning. *arXiv preprint arXiv:2506.01347*, 2025a.

Xudong Zhu, Jiachen Jiang, Mohammad Mahdi Khalili, and Zhihui Zhu. From emergence to control: Probing and modulating self-reflection in language models. *arXiv preprint arXiv:2506.12217*, 2025b.

A Usage of LLMs

Throughout the preparation of this manuscript, Large Language Models (LLMs) were utilized as a writing and editing tool. Specifically, we employed LLMs to improve the clarity and readability of the text, refine sentence structures, and correct grammatical errors. All final content, including the core scientific claims, experimental design, and conclusions, was conceived and written by us, and we take full responsibility for the final version of this paper.

B Related work

Entropy regularization (Mnih et al., 2016; Haarnoja et al., 2018), an early line of work in traditional RL, may mislead actions at critical states (Zhang et al., 2025b) and has been shown to be highly sensitive to the coefficient in LLM training (Cheng et al., 2025; Cui et al., 2025b). (Liu et al., 2025) argues that the KL penalty preserves entropy and acts as a regularizer, ensuring that the online policy remains close to a stable reference, which stabilizes learning and reduces overfitting to misleading reward signals. Nevertheless, the KL divergence term between the current policy π_θ and the reference policy π_{ref} in the original form (Shao et al., 2024) is excluded in our work, since its practical impact is often negligible or counterproductive for reasoning tasks, as demonstrated in recent works (Yu et al., 2025; Chu et al., 2025; Hu et al., 2025). One typical approach to address entropy collapse is by raising the sampling temperature during inference. However, recent findings in (Luo et al., 2025) suggest that while this method postpones the onset of entropy collapse, it does not prevent it, as entropy continues to decrease progressively throughout the training process. Recent studies have sought to mitigate entropy collapse by adjusting key elements of policy optimization, such as PPO-style ratio clipping (Yu et al., 2025; Yang et al., 2025b), balancing positive and negative samples (Zhu et al., 2025a), and applying KL regularization (Liu et al., 2025). However, these methods are broad and lack fine-grained control at the token level, with their mechanisms often not fully explained in a unified or principled way. Several methods attempt to encourage exploration via an entropy-induced advantage (Cheng et al., 2025; Tan & Pan, 2025; Wang et al., 2025c;b; Deng et al., 2025). In practice, however, we found this design often fails to reliably mitigate entropy collapse because it disproportionately strengthens learning on high-entropy tokens and thereby magnifies entropy change, leading to unreliable entropy control. Although prior work (Cui et al., 2025b) considers entropy change, the resulting estimation is distorted (see Figure 1) due to its unreasonable state-equivalence assumption. Notably, its entropy-control scheme (i) enforces a hard binary split by entropy change without considering their intra-group differentiation, and (ii) may hinder the learning process, since high-entropy-change tokens that are informative for exploration are over-penalized.

To summarize existing methods more clearly, we reformulate them through the lens of token-level gradients.

A Token-level Gradient Reweighting Perspective for Shaping Policy Entropy: Existing entropy intervention methods can be unified into a gradient reweighting framework and subsequently examined their respective impacts on policy entropy.

The policy gradient of off-policy optimization can be expressed as follows:

$$\nabla_\theta J(\theta) = \mathbb{E}_{q\sim\mathcal{D},\ \{o_i\}\sim\pi_{\text{old}}(\cdot|q)}\left[\frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|} w_{i,t}(q)\nabla_\theta \log \pi_\theta(o_{i,t}\mid q, o_{i,<t})\right]. \tag{11}$$

For GRPO in Eq. (2), $w_{i,t}(q) = \mathbb{I}_{\text{clip}}\, r_{i,t}A_{i,t}$, where

$$\mathbb{I}_{\text{clip}} = \begin{cases} 0, & A_{i,t} > 0 \ \text{ and } \ r_{i,t} > 1+\varepsilon, \\ 0, & A_{i,t} < 0 \ \text{ and } \ r_{i,t} < 1-\varepsilon, \\ 1, & \text{otherwise}, \end{cases} \tag{12}$$

where $r_{i,t} = \frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\text{old}}(o_{i,t}|q,o_{i,<t})}$ denotes the importance sampling ratio. Advantage $A_{i,t}$ is calculated by reward $R_{i,t}$. For brevity and uniformity, let

$$w_{i,t}(q) = \mathbb{I}_{\text{clip}} r_{i,t}A_{i,t} + \beta\mathcal{R}\left(\pi_\theta\right) \tag{13}$$

, where $\mathcal{R}\left(\pi_\theta\right)$ is the regularization. Table 3 briefly summarizes existing methods based on their interventions on token-level weight $w_{i,t}(q)$. It is evident that existing methods can be categorized into different token-level gradient reweighting schemes, depending on factors such as advantage $A_{i,t}$, generation probability $\pi_\theta(o_{i,t} \mid q, o_{i,<t})$, conditional entropy $\mathcal{H}_{i,t}$, etc. **These methods can be broadly summarized as increasing or suppressing the training weights of tokens that satisfy certain properties.** Our analysis explains why these methods are effective or not. Our proposed STEER adopts reweighting based on token-level entropy change, which is more fundamental for entropy control.

Method	Intervention
DAPO / DCPO (Yu et al., 2025) (Yang et al., 2025b)	$\mathbb{I}_{\text{clip}} = \begin{cases} 0, & A_{i,t} > 0 \text{ and } r_{i,t} > 1 + \varepsilon_{\text{high}}, \\ 0, & A_{i,t} < 0 \text{ and } r_{i,t} < 1 - \varepsilon_{\text{low}}, \\ 1, & \text{otherwise} \end{cases}$
KL penalty (Shao et al., 2024)	$\mathcal{R}\left(\pi_\theta\right) = \frac{\pi_{\text{ref}}(o_{i,t} \mid q, o_{i,<t})}{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}$
Entropy Regularization (He et al., 2025b)	$\mathcal{R}\left(\pi_\theta\right) = -\log \pi_\theta(o_{i,t} \mid q, o_{i,<t})$
Unlikeliness (He et al., 2025a)	$\hat{R}_{i,t} = R_{i,t}\left(1 - \beta_{\text{rank}} \frac{G - \text{rank}(o_i)}{G}\right), \ \beta_{\text{rank}} > 0$
W-REINFORCE (Zhu et al., 2025a)	$\hat{A}_{i,t} = \begin{cases} \lambda, & A_{i,t} > 0 \\ 1, & A_{i,t} < 0 \end{cases}, \quad \lambda < 1$
Entropy Advantage (Cheng et al., 2025)	$\hat{A}_{i,t} = A_{i,t} + \min\left(\alpha \cdot \mathcal{H}_{i,t}^{\text{detach}}, \frac{\lvert A_{i,t} \rvert}{\kappa}\right), \quad \alpha > 0, \kappa > 1$
GTPO (Tan & Pan, 2025)	$\hat{R}_{i,t} = R_{i,t} + \alpha \frac{\mathcal{H}_{i,t}}{\frac{1}{d_t}\sum_{k=1}^{d_t} \mathcal{H}_{k,t}}$, for $R_{i,t} > 0$
EDGE-GRPO (Zhang et al., 2025c)	$\hat{A}_i = \frac{A_i}{\hat{\mathcal{H}}_i}, \ \hat{\mathcal{H}}_i$: normalized entropy across the group
PPL-based (Deng et al., 2025)	$\hat{A}_{i,t} = A_{i,t}(1 - \alpha \text{log-PPL}(o_i)), \ \alpha > 0$
Position-based (Deng et al., 2025)	$\hat{A}_{i,t} = A_{i,t} + \gamma \text{sign}(A_{i,t}) \sigma(r_{it}) \quad r_{it}$: token's relative position
Forking Tokens (Wang et al., 2025c)	$\mathbb{I}_{\text{clip}} = \mathbb{I}_{\text{clip}} \land \mathbb{I}(\mathcal{H}_{i,t} > \tau_{\mathcal{B}}), \ \tau_{\mathcal{B}}$: threshold in batch $\mathcal{B}$

Table 3: A Token-level Gradient Reweighting Perspective for Shaping Policy Entropy.

Besides, recent studies (Wang et al., 2025c;a) highlight the importance of high-entropy tokens for reasoning and propose various mechanisms to strengthen their training. This does not conflict with our approach STEER as STEER explicitly controls token entropy changes to avert training collapse while preserving learning on critical tokens.

C Theorem Proof Details

Theorem 1. *(First–order entropy change) Let the policy model π_θ follows Assumption 1. The change of conditional entropy between two update steps is defined as $\Delta\mathcal{H}_{it} \triangleq \mathcal{H}(\pi_\theta^{k+1} \mid s_{i,t}) - \mathcal{H}(\pi_\theta^{k} \mid s_{i,t})$. Then the first-order estimation of $\Delta\mathcal{H}_{it}$ in Eq. 2 is*

$$\Omega_{i,t} = -\eta \, \mathbb{E}_{a \sim \pi_\theta^k(\cdot|s_{i,t})} \, w_{i,t}(1 - \pi_\theta^k(a|s_{i,t}))^2 \left(\log \pi_\theta^k(a|s_{i,t}) + \mathcal{H}(\pi_\theta^k \mid s_{i,t})\right), \tag{14}$$

where η is the learning rate, $w_{i,t} = \mathbb{I}_{clip} \, r_{i,t} \, A_{i,t}$ is per-token weight.

Proof. The proof is similar to that of (Liu, 2025). Taking the first-order Taylor expansion, we have

$$\begin{aligned}\Delta\mathcal{H}_{it} &\triangleq \mathcal{H}(\pi_\theta^{k+1} \mid s_{i,t}) - \mathcal{H}(\pi_\theta^{k} \mid s_{i,t}) \\ &\approx \left\langle \nabla_\theta \mathcal{H}\left(\pi_\theta^k \mid s_{i,t}\right), z^{k+1} - z^k \right\rangle .\end{aligned}$$

Since we have the log trick $\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}[\nabla_\theta \log \pi_\theta(a \mid s)] = 0$, the gradient term can be derived as

$$\begin{aligned}\nabla_\theta\mathcal{H}(\pi_\theta \mid s) &= \nabla_\theta\mathcal{H}(\pi_\theta(\cdot \mid s)) \\ &= \nabla_\theta\big(-\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\big[\log \pi_\theta(a \mid s)\big]\big) \\ &= -\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}[\nabla_\theta \log \pi_\theta(a \mid s) + \log \pi_\theta(a \mid s)\, \nabla_\theta \log \pi_\theta(a \mid s)] \\ &= -\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}[\log \pi_\theta(a \mid s)\, \nabla_\theta \log \pi_\theta(a \mid s)] .\end{aligned}$$

Then we have

$$\begin{aligned}\Delta\mathcal{H}_{it} =& \big\langle \nabla_\theta\mathcal{H}(\theta^k \mid s_{i,t}),\, (z^{k+1} - z^k)\big\rangle \\ =& - \Big\langle \mathbb{E}_{a\sim\pi_\theta^k(\cdot|s_{i,t})}\big[\, \log \pi_\theta(a \mid s_{i,t})\, \nabla_\theta \log \pi_\theta(a \mid s_{i,t})\,\big],\, \theta^{k+1} - \theta^k \Big\rangle \\ =& - \mathbb{E}_{a\sim\pi_\theta^k(\cdot|s_{i,t})}\Big[\, \log \pi_\theta(a \mid s_{i,t})\, \big\langle \nabla_\theta \log \pi_\theta(a \mid s_{i,t}),\, \theta^{k+1} - \theta^k \big\rangle\Big] \\ =& - \mathbb{E}_{a\sim\pi_\theta^k(\cdot|s_{i,t})}\left[\log \pi_\theta(a \mid s_{i,t}) \sum_{a'\in\mathcal{A}} \frac{\partial \log \pi_\theta(a \mid s_{i,t})}{\partial \theta_{s_{i,t},a'}} \left(\theta^{k+1}_{s_{i,t},a'} - \theta^{k}_{s_{i,t},a'}\right)\right] \\ =& - \mathbb{E}_{a\sim\pi_\theta^k(\cdot|s_{i,t})}\left[\log \pi_\theta(a \mid s_{i,t}) \sum_{a'\in\mathcal{A}} \big(\mathbf{1}\{a = a'\} - \pi(a' \mid s_{i,t})\big) \left(\theta^{k+1}_{s_{i,t},a'} - \theta^{k}_{s_{i,t},a'}\right)\right] \\ =& - \mathbb{E}_{a\sim\pi_\theta^k(\cdot|s_{i,t})}\Bigg[\Big(\, \log \pi_\theta(a \mid s_{i,t}) - \mathbb{E}_{\hat{a}\sim\pi_\theta^k(\cdot|s_{i,t})} \log \pi_\theta(a \mid s_{i,t})\Big) \\ & \Big(\theta^{k+1}_{s_{i,t},a} - \theta^{k}_{s_{i,t},a} - \mathbb{E}_{a'\sim\pi_\theta^k(\cdot|s_{i,t})}\big(\theta^{k+1}_{s_{i,t},a'} - \theta^{k}_{s_{i,t},a'}\big)\Big)\Bigg] \\ =& - \mathbb{E}_{a\sim\pi_\theta^k(\cdot|s)}\left[\log \pi_\theta^k(a|s) + \mathcal{H}(\cdot|s)\right]\Big[\big(1 - \pi_\theta^k(a|s)\big)\Big(z^{k+1}_{s_{i,t},a} - z^{k}_{s_{i,t},a}\Big)\Big] \\ =& - \mathbb{E}_{a\sim\pi_\theta^k(\cdot|s)}\left[\log \pi_\theta^k(a|s) + \mathcal{H}(\cdot|s)\right]\Big[w(s|a)\big(1 - \pi_\theta^k(\alpha|s)\big)^2\Big] ,\end{aligned}$$

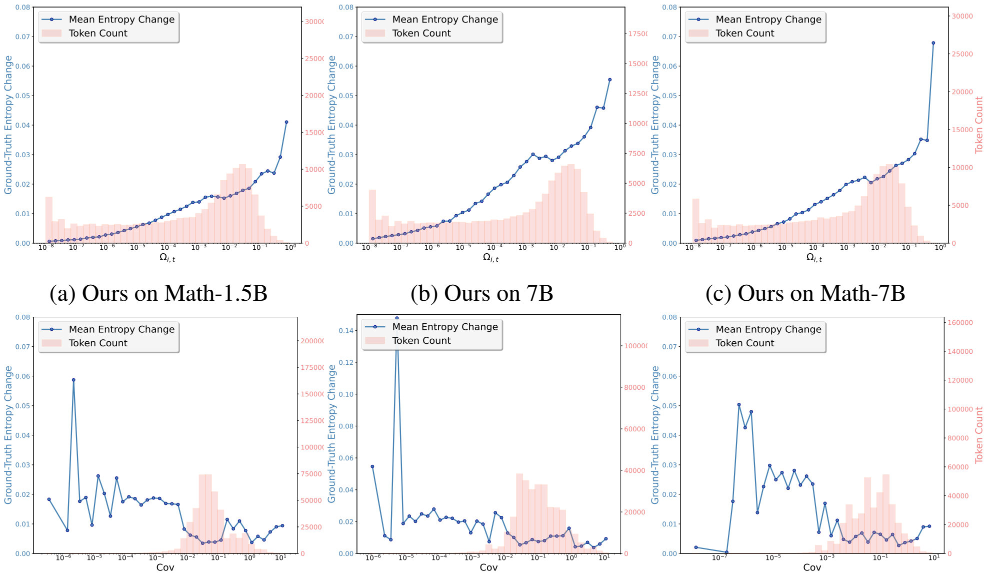
where $w(s|a)$ is the weight in the policy gradient. The derivation from the fourth-to-last equation to the third-to-last equation is due to the zero expectation of the second term. The derivation from the third-to-last equation to the second-to-last equation is due to the parameter-independent softmax: when action a is sampled under state s, only its corresponding parameter $z_{s,a}$ is updated, while the parameters corresponding to other actions $z_{s,a'}$ remain unchanged.

□

D Additional Experiments

D.1 Entropy Change Estimation Comparison

We recorded the token entropy changes for the first 10 training steps across different models and datasets. Figure 18 and 19 show the results on dataset DAPO-Math-17k, while Figure 20 and 21 show the results on dataset Math. It can be observed that our method exhibits a clear positive correlation with ground-truth entropy change, which strongly supports our theoretical framework. By contrast, the estimation scheme in (Cui et al., 2025b) exhibits no clear correlation.



(a) Ours on Math-1.5B

(b) Ours on 7B

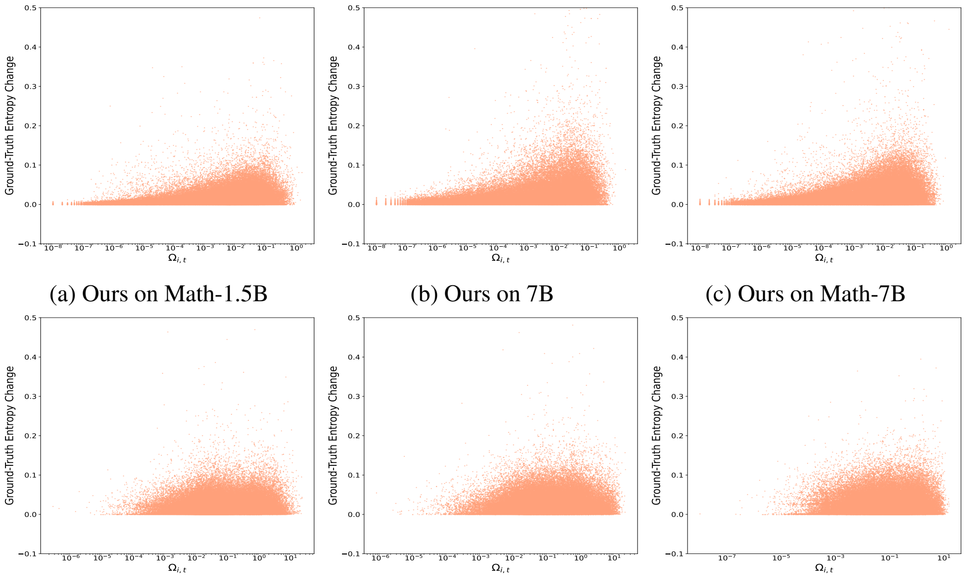
(c) Ours on Math-7B

(d) Cov on Math-7B

(e) Cov on Math-7B

(f) Cov on Math-7B

Figure 18: Entropy Change on DAPO-Math-17k.



(a) Ours on Math-1.5B

(b) Ours on 7B

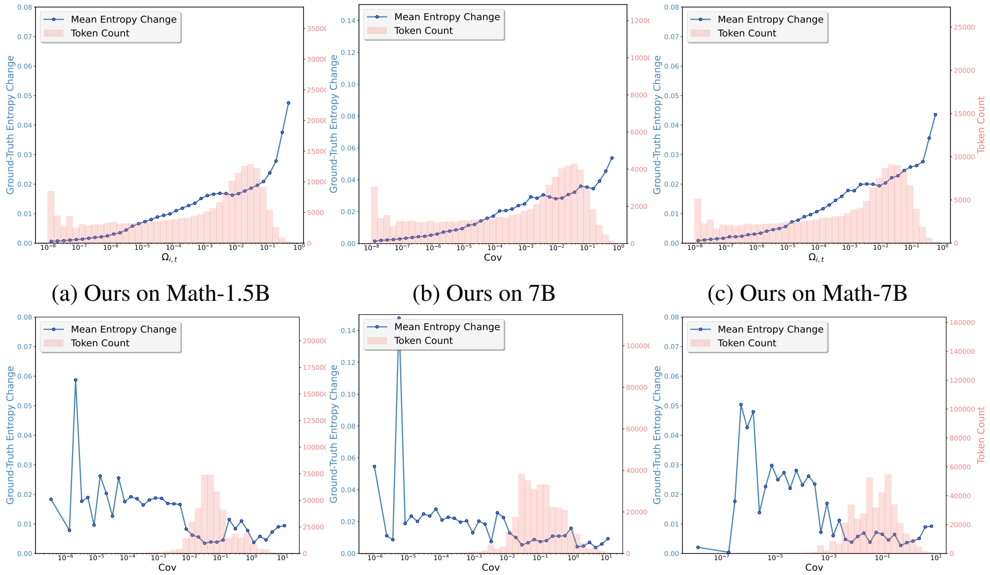
(c) Ours on Math-7B

(d) Cov on Math-1.5B

(e) Cov on 7B

(f) Cov on Math-7B

Figure 19: Entropy Change scatters on DAPO-Math-17k.



(a) Ours on Math-1.5B

(b) Ours on 7B

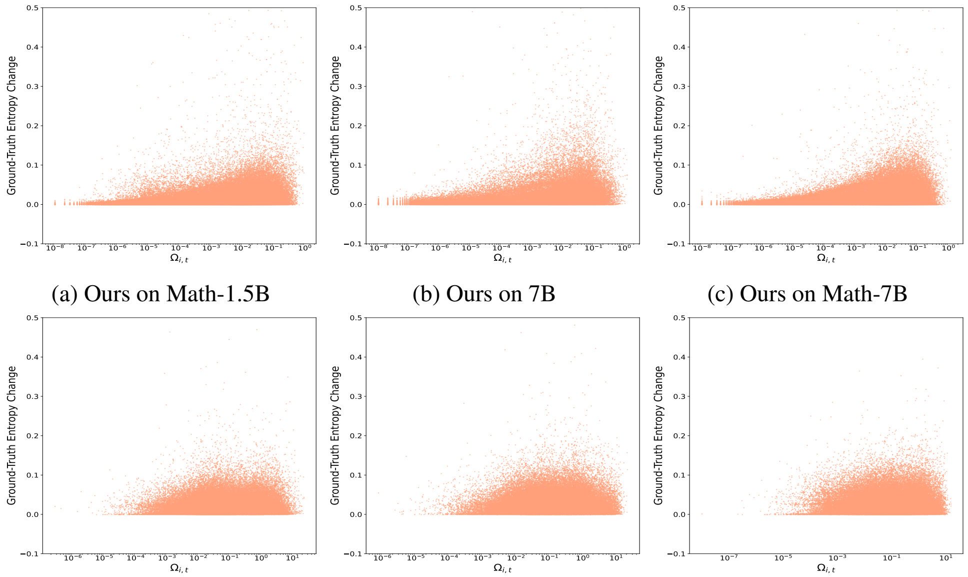
(c) Ours on Math-7B

(d) Cov on Math-7B

(e) Cov on Math-7B

(f) Cov on Math-7B

Figure 20: Entropy Change on Math.



(a) Ours on Math-1.5B

(b) Ours on 7B

(c) Ours on Math-7B

(d) Cov on Math-1.5B

(e) Cov on 7B

(f) Cov on Math-7B

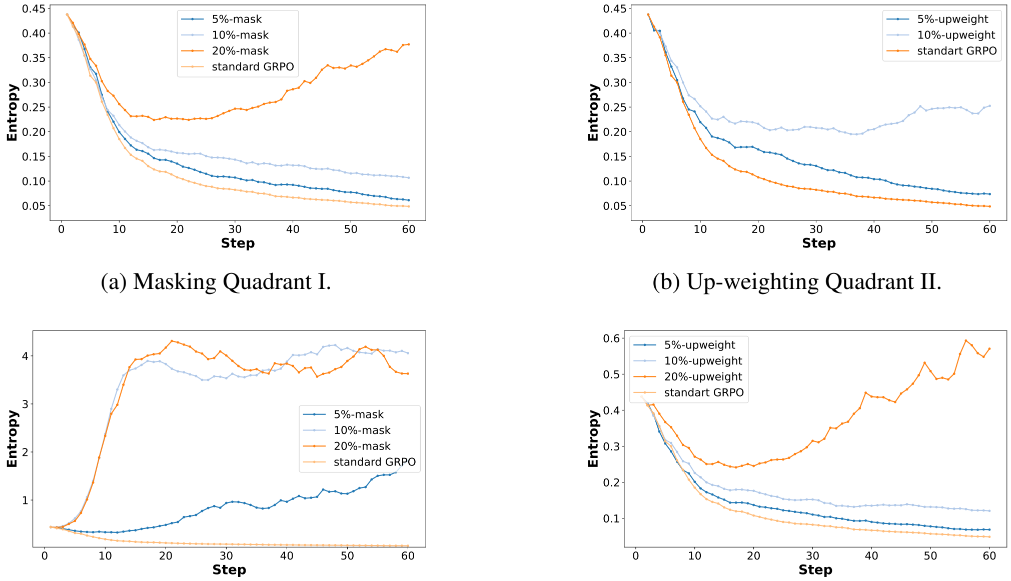
Figure 21: Entropy Change scatters on Math.

D.2 Influencing Entropy Dynamics by Strengthening or Weakening the Quadrants

For experiments in Figure 6, we randomly select samples with a generation probability greater than 0.8 and an advantage greater than 0, as well as those with a generation probability less than 0.2 and an advantage less than 0, and randomly mask 10% of such tokens. Similarly, for samples with a generation probability greater than 0.8 and an advantage less than 0, or a generation probability less than 0.2 and an advantage greater than 0, we set the token weight for 10% of such tokens to twice the original token weight.

To further validate the patterns of entropy change with advantage and probability in Figure 4, we strengthen (up-weighting) or weaken (masking) each of the four quadrants at different intensities to induce entropy increases or decreases, respectively. Unlike the setup in Figure 6 where 10% tokens are intervened, we present a more comprehensive validation here.

Figure 22 shows interventions applied to each quadrant with the goal of increasing entropy, using standard GRPO ($\varepsilon_{\text{high}=0.2}, \varepsilon_{\text{low}=0.2}$) as the baseline; while Figure 23 presents interventions with the goal of decreasing entropy across the four quadrants, using GRPO w/ clip-high ($\varepsilon_{\text{high}=0.28}, \varepsilon_{\text{low}=0.2}$) as the baseline; In each case, the proportion of tokens masked or up-weighted ranges from 5% to 20%. Across all cases, it can be observed that the token-level intervention effects on entropy align with our quantitative analysis framework, and the impact becomes more pronounced as the intervention ratio increases (from 5% to 20%). For example, in Figure 22b, compared to standard GRPO, up-weighting Quadrant II yields a marked increase in policy entropy over standard GRPO (we exclude the 20% up-weight case because it produces excessively high entropy). This indicates that the clip-high mechanism in DAPO (Yu et al., 2025) and unlikeliness He et al. (2025a)can be viewed as a special instance of this intervention. In summary, the overall entropy dynamics arise from the joint contributions of the four quadrants; perturbing any one of them can induce a predictable change in the total entropy from our analysis framework.



(a) Masking Quadrant I.

(b) Up-weighting Quadrant II.

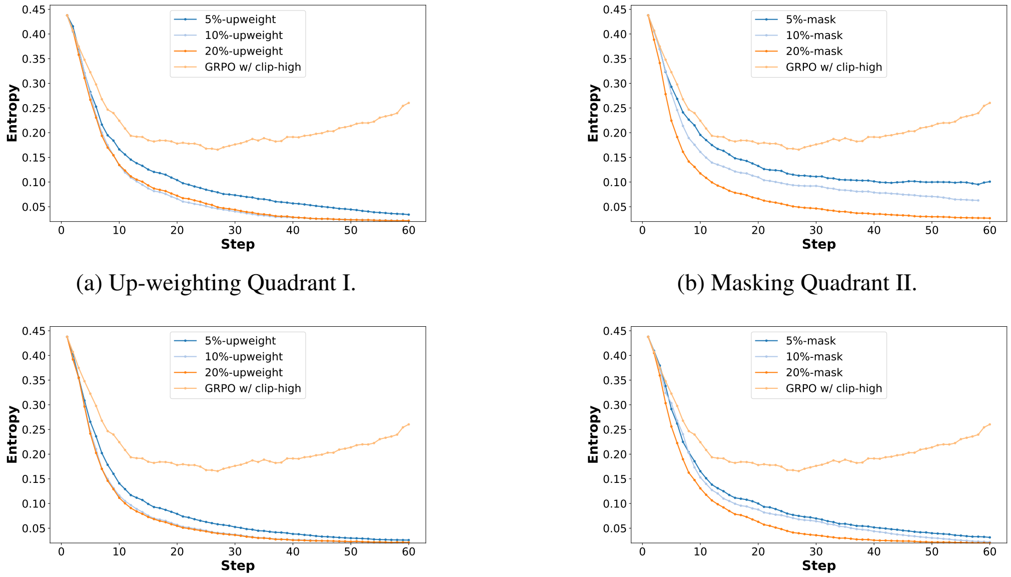
(c) Masking Quadrant III.

(d) Up-weighting Quadrant IV.

Figure 22: Increasing entropy in four cases.

D.3 Performance in Extreme Scenarios

We test entropy intervention methods in uncontrolled training scenarios ($\varepsilon_{\text{high}} = 5$ and $\varepsilon_{\text{low}} = 0.99$), with test set accuracy shown in Table 4. It can be observed that, even in training scenarios where the clipping operation is almost completely removed, STEER maintains relatively stable performance



(a) Up-weighting Quadrant I.

(b) Masking Quadrant II.

(c) Up-weighting Quadrant III.

(d) Masking Quadrant IV.

Figure 23: Decreasing entropy in four cases.

compared to other entropy intervention methods and achieves the highest accuracy across all test sets.

Table 4: Performance on test datasets in extreme scenarios.

Method	AIME24	AIME25	AMC23	MATH500	Minerva	Olympiad	Avg.
GRPO	31.6	12.8	66.7	79.0	39.3	40.1	44.9
Entro. Adv.	34.8	13.4	64.3	77.6	37.6	39.9	44.6
Entro. Loss	32.7	14.7	71.3	79.0	36.8	41.4	46.0
Clip-Cov	30.4	14.0	72.3	79.6	37.1	41.7	45.8
STEER	**36.9**	**16.2**	**76.3**	**81.2**	**39.3**	**42.4**	**48.7**

D.4 Performance Comparison on Different Models

The performance experiments on Qwen2.5-Math-1.5B and Qwen2.5-14B shown in Table 5 are compared with the top3 competitors in Table 1 (i.e., OPO, Clip Cov, and Entro. Adv.). STEER also consistently achieves the highest average performance on both Qwen2.5-Math-1.5B (38.1) and Qwen2.5-14B (45.1), demonstrating its superior capabilities in improving model reasoning.

Table 5: Benchmark results of different methods. We report avg@32 for AIME24, AIME25, and AMC23 and avg@1 for others. All results are presented as percentages.

Method	**AIME24**	**AIME25**	**AMC23**	**MATH500**	**Minerva**	**Olympiad**	**Avg.**
				Qwen2.5-Math-1.5B			
base	4.1	2.1	24.7	29.0	9.2	20.5	14.9
GRPO	16.2	7.6	56.0	74.4	26.1	34.6	35.8
OPO	14.8	9.0	58.2	72.2	26.1	35.9	36.0
Entro. Adv.	15.0	9.1	55.7	70.2	26.8	34.9	35.3
Clip-Cov	14.7	8.4	56.0	72.8	26.4	34.9	35.5
STEER	17.2	9.7	61.3	75.4	28.0	36.9	38.1
				Qwen2.5-14B			
base	3.9	2.6	25.8	52.6	15.4	23.0	20.6
GRPO	17.2	13.2	66.3	80.6	38.0	42.2	42.9
OPO	17.8	12.6	68.2	78.6	37.7	42.6	42.9
Entro. Adv.	14.6	9.8	65.6	78.8	36.5	40.9	41.0
Clip-Cov	14.1	13.6	59.8	78.2	38.6	43.2	41.2
STEER	19.3	14.0	70.3	81.6	39.1	46.3	45.1

E Training Settings

E.1 Detailed Information for test dataset

Table 6: Dataset statistics.

Test Datasets	**#Questions**	**Level**
AIME24	30	Olympiad
AIME25	30	Olympiad
AMC23	40	Intermediate
MATH500	500	Advanced
Minerva	272	Graduate
OlympiadBench	675	Olympiad

E.2 Training Details for our method and baselines.

All algorithms are implemented based on the official GRPO codebase within the VeRL framework. We use a learning rate of 1e-6 without warm-up across all experiments. At each rollout step, we generate 8 answers for each of 512 sampled questions, then split the data into 16 mini-batches and train the policy network for 16 gradient steps. Models are trained for at most 150 rollout steps. Unless otherwise specified, we follow GRPO's default design choices with token-level loss normalization without dynamic sampling and KL regularization. For all models, the maximum input length is 1024 and the maximum output length is 3072. All methods save a checkpoint every 10 steps, and the checkpoint achieving the highest AIME24 accuracy is selected for test. All the experiments were conducted on H20 GPUs.