

VFPO: Variance-aware Negative Feedback Policy Optimization for Mitigating Entropy Collapse via Group Statistics

Bo Zhang¹, Zinan Zeng¹, Juntong Zhou¹, Dading Chong²,
Zhenyu Ding¹, Dengpeng Wu¹, Ning Ding¹✉, Peilin Zhou³✉

¹ Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University,

² Peking University, ³ Hong Kong University of Science and Technology
ding.ning@xjtu.edu.cn, zhoupalin@gmail.com

Abstract

Reinforcement Learning with Verifiable Rewards (RLVR) has significantly advanced the reasoning capabilities of Large Language Models (LLMs), while being prone to *entropy collapse*, a rapid degradation of policy entropy that forces the model into premature convergence and limits its performance ceiling. However, existing solutions for mitigating *entropy collapse* often struggle to reconcile the trade-off between exploration capability and convergence efficiency. To address this, we propose **Variance-aware Negative Feedback Policy Optimization (VFPO)**, a framework that dynamically regulates the policy entropy through the lens of group statistics. We first mathematically prove that policy entropy is intrinsically coupled with the variance and mean of group probabilities for sampled trajectories. Leveraging this insight, VFPO operates via a dual-phase strategy: an initial exploration phase employs a closed-loop negative feedback mechanism via both sample-level and token-level reward modulation to maintain group variance and prevent the mean trajectory probability from saturating; then followed by an annealing phase that relaxes these constraints to facilitate deterministic convergence. Extensive experiments on models of distinct scales (1.5B, 3B, 7B) demonstrate that VFPO yields average improvements of 3.34%, 6.52%, and 6.03%, respectively, over the DAPO baseline, and achieves state-of-the-art performance across diverse mathematical benchmarks, effectively expanding the exploration boundary while achieving superior reasoning accuracy. To facilitate reproducibility, we release the code, data and trained models at the anonymous link <https://github.com/carpe1/VFPO>.

1 Introduction

Reinforcement Learning with Verifiable Rewards (RLVR) has become the key technology for enhancing the complex reasoning capabilities of

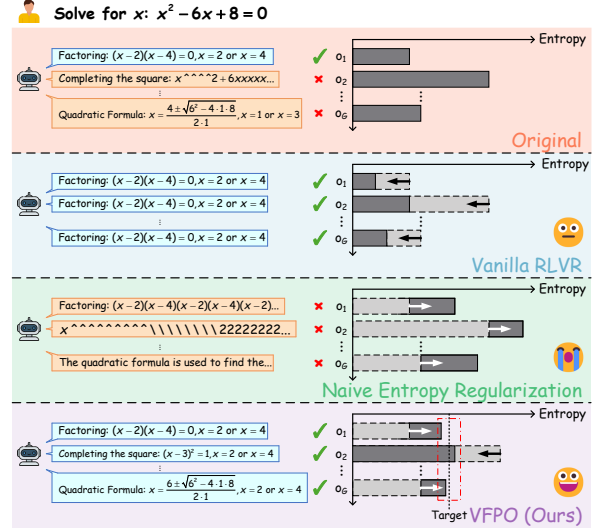


Figure 1: Comparative illustration of policy evolution strategies. **Original:** The pre-trained model generates diverse solutions with mixed correctness. **Vanilla RLVR:** Monotonically decreases entropy, causing the model to prematurely converge to a single solution path. **Naive Entropy Regularization:** Monotonically encourages entropy increase, leading to distribution divergence and incoherent outputs. **VFPO (Ours):** Sets a dynamic target for group statistics. It encourages exploration when variance is low and exploitation when variance is high, successfully identifying multiple diverse correct reasoning paths.

Large Language Models (LLMs) (Guo et al., 2025), yet it suffers from a fundamental *exploration/exploitation* dilemma during fine-tuning. In vanilla RLVR methods such as Group Relative Policy Optimization (GRPO) (Shao et al., 2024), the group-relative advantage concentrates probability mass on the few correct samples in early training, yielding a high intra-group mean and low variance. As shown in the second panel of Figure 1, this drives a monotonic decrease in entropy and locks the policy into a single solution path, a phenomenon known as *entropy collapse*, which causes premature loss of exploration (Yue et al., 2025) and

performance stagnation. Subsequent variants such as DAPO (Yu et al., 2025) improve stability but do not explicitly prevent this early-stage collapse.

A natural remedy is to add an entropy bonus to the objective, as commonly done in policy gradient methods (Williams, 1992; Schulman et al., 2017; Haarnoja et al., 2018). However, in RLVR this naive regularization tends to induce a monotonic entropy increase, driving the policy toward incoherent or incorrect outputs, a failure mode we term *model collapse* (third panel of Figure 1).

To resolve this dilemma, we first prove that policy entropy is proportional to the variance and inversely proportional to the mean of intra-group trajectory probabilities. Building on this insight, we propose **Variance-aware Negative Feedback Policy Optimization (VFPO)**, which replaces monotonic adjustment with a negative feedback loop that drives group statistics toward a dynamic target: entropy is injected when the policy is over-confident and suppressed when it is over-dispersed. The feedback is applied jointly at the *sample* and *token* levels for precise control over critical decision points, and is deactivated during a subsequent *Annealing Phase* to allow natural convergence once sufficient advantageous paths have been mined. Our main contributions are as follows:

1. Through theoretical analysis, we demonstrate that policy entropy is negatively correlated with group probability and positively correlated with group variance.
2. We innovatively propose the negative feedback mechanism **VFPO**, a dual regulation at both the *sample level* and the *token level*, and a phased strategy to maintain the policy entropy within a reasonable range.
3. Empirical evaluations on the Qwen2.5 model family across multiple mathematical reasoning benchmarks demonstrate that VFPO significantly outperforms strong baselines.

2 Related Work

2.1 Reinforcement Learning with Verifiable Rewards

While RLHF (Gao et al., 2023; Ouyang et al., 2022) has achieved remarkable success in aligning LLMs, applying it to complex reasoning tasks, such as mathematical problem solving (Cobbe et al., 2021), remains challenging due to the sparsity and unreliability of reward signals. To address this, the

paradigm of RLVR has emerged as a robust alternative. RLVR leverages the inherent verifiability of specific domains (e.g., formal proofs or code execution) to construct precise, dense, and unbiased reward signals, thereby guiding the policy toward more effective reasoning paths.

A pivotal advancement in this domain is GRPO (Shao et al., 2024). By estimating relative advantages through the comparison of multiple sampled trajectories for a given problem, GRPO eliminates the dependency of value models, resulting in more robust gradient estimation. The efficacy of this approach has been further validated by the success of DeepSeek-R1 (Guo et al., 2025), sparking significant interest in the RLVR paradigm. Building upon this foundation, recent works have focused on enhancing training stability and efficiency. For instance, DAPO (Yu et al., 2025) integrates techniques such as Clip-Higher, dynamic sampling, and token-level policy gradient loss with extended reward shaping to stabilize the training process. Concurrently, Dr.GRPO (Liu et al., 2025) addresses the inductive biases related to response length and problem difficulty inherent in the original GRPO framework. Collectively, these methods establish RLVR as a scalable and verifiable pathway for advancing the logical reasoning capabilities of LLMs.

2.2 Entropy Collapse in Large Language Models

Despite the promise of RLVR, LLMs often suffer from *entropy collapse*, where the policy distribution prematurely converges, eroding exploration and generalization during RL training.

Recent work tackles this from several angles. Cui et al. (2025) attribute entropy variance to the covariance between action probabilities and logits, and propose Clip-Cov and KL-Cov to suppress high-covariance tokens. EMPG (Wang et al., 2025b) rescales the update step by the current entropy to improve long-horizon coherence, while SIREN (Jiang et al., 2025) applies top- p and peak-entropy masking to balance exploration. DisCO (Li et al., 2025) replaces GRPO’s group-relative objective with a discriminative one under a KL constraint, stabilizing entropy without clipping. From a signal-decomposition view, Zhu et al. (2025) find that upweighting negative-sample reinforcement alone stabilizes training and lifts Pass@ k , and QAE (Wu et al., 2026) swaps the group mean baseline for a K -quantile to guard against both entropy collapse and explosion.

3 Problem Formulation

We conceptualize "entropy collapse" in LLM not merely as a scalar metric decay, but as a coupled distributional collapse mode characterized by the **vanishing of sample group variance** and the **saturation of group means**. Mechanistically, the rapid reduction of group variance constrains the search space, effectively pruning the "long-tail" exploration paths required to discover complex, high-reward solutions that initially possess low likelihood. Simultaneously, the saturation of the group mean drives the policy distribution toward a state of excessive confidence and rigidity; this lack of stochasticity eliminates the necessary diversity for effective advantage propagation, inevitably trapping the model in suboptimal local optima. For clarity, a summary of the mathematical notations throughout this work is provided in Appendix A.

3.1 Inverse Monotonicity of Group Mean

Let $\mathcal{Y} = \{y^{(1)}, \dots, y^{(|G|)}\}$ be a batch of independent responses sampled from policy π_θ . For a sample $y^{(i)}$ of length n , where $\pi_{i,j}$ denotes the predicted probability of its j -th token, we define the sample's geometric mean probability as $\pi_i = \left(\prod_{j=1}^n \pi_{i,j}\right)^{\frac{1}{n}}$. This metric reflects the average confidence of the model in generating the sequence. Correspondingly, we define the normalized log-likelihood L_i directly via π_i :

$$L_i = \ln \pi_i = \frac{1}{n} \sum_{j=1}^n \ln \pi_{i,j} \quad (1)$$

We define the group statistics (Mean μ_{group} , Variance σ_{group}^2) over the entire batch:

$$\begin{aligned} \mu_{\text{group}} &= \frac{1}{|G|} \sum_{i=1}^{|G|} \pi_i, \\ \sigma_{\text{group}}^2 &= \frac{1}{|G|} \sum_{i=1}^{|G|} (\pi_i - \mu_{\text{group}})^2 \end{aligned} \quad (2)$$

We first establish the analytical relationship between the Group Mean μ_{group} and the system Entropy H . We aim to prove that a reduction in entropy mathematically necessitates an increase in the expected geometric probability.

Theorem 3.1 (Inverse Monotonicity of Group Mean). *Let $\mu_{\text{group}} = \mathbb{E}[\exp(L)]$ be the group mean of the geometric probabilities and the policy entropy $H = -\mathbb{E}[L]$, the sensitivity of the mean with*

respect to the entropy is negative: $\frac{\partial \mu_{\text{group}}}{\partial H} < 0$. (Detailed proof in the Appendix B.1)

This theorem confirms that the partial derivative of the group mean with respect to entropy is less than zero, which means that the group mean is inversely correlated to entropy.

3.2 Positive Correlation of Variance

In this section, we rigorously analyze the functional dependence of the Group Variance on the policy entropy. We explicitly model the group variance as the population variance of the geometric mean probabilities and establish a strict positive correlation via singular perturbation analysis on the probability manifold.

Lemma 3.2 (Variance Bound for Bounded Variables). *For any random variable X with support $X \in [a, b]$, the variance is bounded by the distance of the mean from the endpoints:*

$$\text{Var}(X) \leq (\mu_X - a)(b - \mu_X) \quad (3)$$

This is a direct consequence of the Bhatia-Davis inequality.

Theorem 3.3 (Positive Correlation of Variance in Convergence Regime). *Let π_θ be a policy in the convergence regime, defined as the state where $\mu_{\text{group}} > 0.5$. As the policy entropy H decreases, the upper bound of Group Variance σ_{group}^2 decreases, approaching zero when $H \rightarrow 0$. (Proof in the Appendix B.2)*

For convenience, we denote the geometric mean probability π_i as the random variable p . To quantify the variance propagation from log-space to probability-space, we apply the First-Order Delta Method around $\mu_L = -H(\pi)$:

$$\begin{aligned} \text{Var}[p] &\approx [(e^L)'_{L=-H(\pi)}]^2 \text{Var}[L] \\ &= e^{-2H(\pi)} \frac{V(\pi)}{n^2}. \end{aligned} \quad (4)$$

Equation (4) reveals a competition between the exponential growth of $e^{-2H(\pi)}$ and the collapse of Varentropy $V(\pi)$ as entropy decreases. To resolve this ambiguity, we provide a rigorous singular perturbation analysis in Appendix B.3, proving that under the "Core-Tail" regime governing LLM generation, the variance is strictly positively correlated with entropy ($\partial \text{Var}[p] / \partial H > 0$).

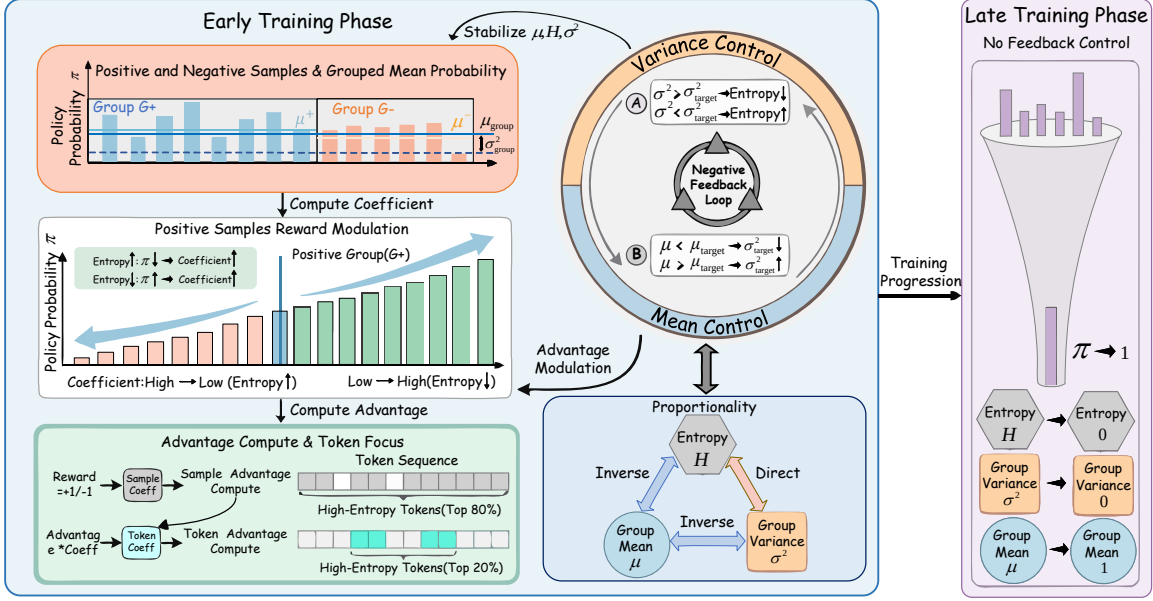


Figure 2: Schematic overview of the Variance-aware Negative Feedback Policy Optimization (VFPO) framework. **Left (Early Phase):** A statistical regulation system monitors the group variance σ^2 and group mean μ . A negative feedback loop dynamically adjusts the reward coefficients to maintain entropy within an optimal exploration corridor. **Right (Late Phase):** The feedback constraints are relaxed ($\sigma_{\text{target}}^2 \rightarrow 0$) to anneal the policy towards a low-entropy, high-confidence state.

4 Methodology

To mitigate entropy collapse and sustain stable exploration, we propose the Variance-aware Negative Feedback Policy Optimization (VFPO) framework, which dynamically modulates the reward structure at both the sample and token levels to directly regulate the variance and mean of group probabilities for sampled trajectories. Leveraging the intrinsic mathematical coupling proven in Section 3, this explicit regulation allows us to *indirectly* govern the policy entropy. The overall workflow is illustrated in Figure 2: during the early training phase, VFPO maintains entropy stability by dynamically switching between entropy expansion and reduction modes via both sample-level and token-level reward modulation; in the late training phase, this control is relaxed to facilitate convergence. The detailed pseudocode is provided in Algorithm 1 in Appendix C.5. In the following subsections, we elaborate on *Sample-Level Reward Modulation* and *Token-Level Granular Control*, the two core innovations that sustain effective exploration during the early training phase.

4.1 Sample-Level Reward Modulation

We use a modulated reward $R_i = c_i \cdot r_i$, where the discrete reward $r_i \in \{1, -1\}$ is scaled by a scalar

c_i . c_i comprises a base incentive c_{base} and a slope correction coefficient α :

$$c_i = c_{\text{base}} + \alpha \cdot (\pi_i - \mu_{\text{ref}}) \quad (5)$$

Here, $\mu_{\text{ref}} \in \{\mu^+, \mu^-\}$ denotes the conditional mean probability for the correct (G^+) or incorrect (G^-) subsets, formally defined as $\mu^\pm = \mathbb{E}_{i \in G^\pm} [\pi_i]$.

4.1.1 Group Mean-Variance Coupled Regulation

To sustain stable exploration during the early training phase, we introduce a Variance-aware Negative Feedback Loop. This framework utilizes the variance target σ_{target}^2 as a pivot to explicitly trigger mode switches, thereby dynamically correcting variance deviations.

i) Dual-Mode Switching Logic: As illustrated in Figure 2, the modulation strategy is orchestrated by the variance deviation $\sigma_\Delta^2 = \sigma_{\text{group}}^2 - \sigma_{\text{target}}^2$. This deviation serves as a control signal, dictating whether the system should encourage entropy growth or reduction. By dynamically switching between these two regimes, VFPO effectively maintains entropy stability during the early training phase. We formally define these distinct opera-

tional modes as follows:

$$c_{\text{base}} = \begin{cases} 1 + acc_{\text{group}}, & \sigma_{\Delta}^2 > 0 \\ 1 + |\sigma_{\Delta}^2|, & \sigma_{\Delta}^2 < 0 \end{cases} \quad (6)$$

acc_{group} represents the accuracy rate of the entire sampled batch, which serves to naturally assign higher reward values to high-performing batches when $\sigma_{\Delta}^2 > 0$.

Entropy Reduction Mode ($\sigma_{\Delta}^2 > 0$): Since σ_{group}^2 exceeds the target, we set α to a positive value to reward high-probability samples, effectively suppressing entropy. **Entropy Expansion Mode** ($\sigma_{\Delta}^2 < 0$): Since σ_{group}^2 is below the target, we set α to the opposite of its original value to reward low-probability correct samples, effectively injecting entropy.

ii) Dynamic Target Coupling: As illustrated in Figure 2, we establish a dynamic coupling where the target variance σ_{target}^2 correlates with the group mean μ_{group} . High probability ($\mu_{\text{group}} \uparrow$) elevates the target to induce **Entropy Expansion Mode** for diversity, whereas low probability ($\mu_{\text{group}} \downarrow$) lowers the target to facilitate **Entropy Reduction Mode** for stability. The coupling relationship is defined as:

$$\sigma_{\text{target}}^2 = \sigma_{\text{base}}^2 \cdot \left(1 + \frac{\mu_{\text{group}} - \mu_{\text{target}}}{1 - \mu_{\text{target}}} \right) \quad (7)$$

σ_{base}^2 serves as a baseline value to facilitate the dynamic calculation of σ_{target}^2 , allowing it to adjust relative to a stable reference. This coupling ensures policy dynamically demands higher diversity as the sample probability increases, preventing entropy collapse and sustaining exploration capability.

4.1.2 Low-probability Attenuation

For error samples ($y^{(i)} \in G^-$), the reference mean is assigned as $\mu_{\text{ref}} = \mu^-$. To sustain the model’s exploratory potential, a penalty attenuation mechanism is implemented, which remains active across both Concentration and Expansion modes. This modulation is defined as:

$$c_i = 1 + \alpha \cdot \min(0, \pi_i - \mu^-) \quad (8)$$

As π_i decreases further below the group mean μ^- , the coefficient c_i is reduced more significantly (yielding a smaller negative reward).

4.2 Token-Level Granular Control

4.2.1 Sign-Preserving Reward Standardization

The modulated rewards R_i require normalization to be transformed into advantage values A_i . However,

Since the modulated rewards are continuous rather than binary, standard Z-score normalization (Shao et al., 2024) may cause sign flipping (e.g., a small positive reward becoming negative), which corrupts the semantic meaning of "correctness". We propose a Sign-Preserving Standardization method. Let $\mathcal{R} = \{R_i\}$ be the set of modulated rewards with mean $\bar{R}$.

- **Case $\bar{R} > 0$:** Identify the minimum positive reward R_{min}^+ . If $R_{\text{min}}^+ < \bar{R}$, shifting by the mean would flip its sign. We instead shift by R_{min}^+ (i.e., $R_i \leftarrow R_i - R_{\text{min}}^+$). To maintain zero mean, negative rewards are scaled proportionally.

- **Case $\bar{R} < 0$:** A symmetric operation is applied using the maximum negative reward to prevent negative-to-positive flipping. Finally, Similar to the Z-score method, the modulated rewards are divided by their standard deviation. Through this sign-preserving standardization, we transform the modulated rewards R_i into advantage values A_i .

4.2.2 Token Advantage Modulation

To achieve finer control, we modulate the advantage values at the token level based on the global variance state. We define a token-level scaling coefficient $c_{i,t}$ that adapts to the probability distribution of the current policy.

For **correct samples**, we amplify the advantage of low-probability tokens to encourage exploration; otherwise, we reinforce high-probability tokens:

$$c_{i,t}^+ = \begin{cases} \sqrt{\frac{\mu^+}{\pi_{i,t}}}, & \text{if } \sigma_{\text{group}}^2 < \sigma_{\text{target}}^2 \\ \sqrt{\frac{\pi_{i,t}}{\mu^+}}, & \text{otherwise} \end{cases} \quad (9)$$

For **incorrect samples**, we apply a symmetric logic using the negative group mean μ^- to regulate error suppression:

$$c_{i,t}^- = \begin{cases} \sqrt{\frac{\pi_{i,t}}{\mu^-}}, & \text{if } \sigma_{\text{group}}^2 < \sigma_{\text{target}}^2 \\ \sqrt{\frac{\mu^-}{\pi_{i,t}}}, & \text{otherwise} \end{cases} \quad (10)$$

We apply this modulation selectively using a *High-Entropy Mask* $m_{i,t}^{\text{entropy}}$, which is 1 if the token’s entropy $H_{i,t}$ is in the top- K percentile of the sequence, and 0 otherwise. The final modulated advantage $\tilde{A}_{i,t}$ is:

$$\tilde{A}_{i,t} = \text{clip}(C_{i,t}^{\pm}, 1 - \tilde{\epsilon}_{\text{low}}, 1 + \tilde{\epsilon}_{\text{high}}) A_{i,t} m_{i,t}^{\text{entropy}} \quad (11)$$

This ensures that entropy regulation is targeted at critical branching points in the generation, rather than diffused across trivial tokens.

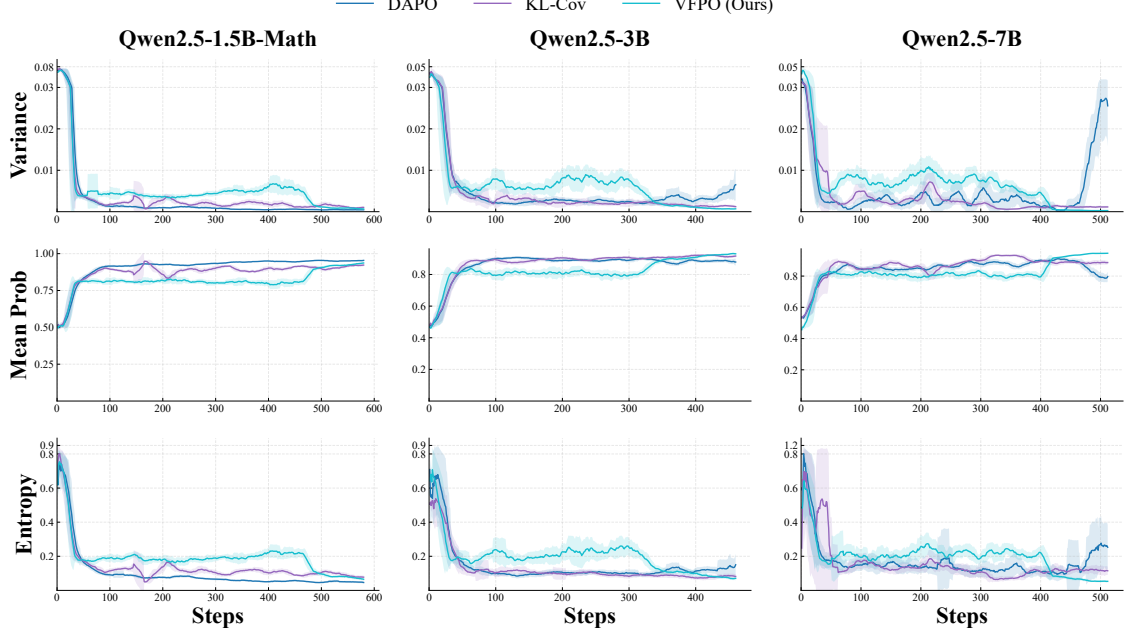


Figure 3: Evolution of group statistics (Group Variance, Mean Probability, and Entropy) throughout the training process across three model scales. Shaded regions indicate standard deviation. VFPO prevents the premature distributional collapse observed in baselines by maintaining a high-variance during the initial exploration phase.

4.2.3 Difference-Based Clipping

Standard PPO clips the probability ratio $r_t(\theta) = \pi_{\text{new}}/\pi_{\text{old}}$, which can be unfair to tokens with different absolute probability scales. To address this, we introduce a Difference-Based Clipping mechanism. We constrain the absolute change in probability:

$$\Delta\pi = \text{clip}(\pi_\theta - \pi_{\theta_{\text{old}}}, -\epsilon_{\text{low}}, \epsilon_{\text{high}}) \quad (12)$$

The policy update ratio is then reconstructed as:

$$r_{i,t}(\theta) = \frac{\pi_{\theta_{\text{old}}} + \Delta\pi}{\pi_{\theta_{\text{old}}}} \quad (13)$$

Our empirical analysis in Figure 5 validates that this mechanism, by permitting a broader clipping range, elevates the clipping activation probability to a more reasonable and effective level. Unlike ratio-based clipping, our approach maintains sufficient gradient magnitude to "activate" low-probability tokens effectively. The rigorous mathematical derivation can be found in the Appendix B.4.

5 Experiments

In this section, we conduct a comprehensive empirical evaluation of the proposed VFPO framework. To demonstrate VFPO’s robustness and scalability, we validate our method across a wide range of model scales (from 1.5B to 7B) and a diverse set of challenging benchmarks.

5.1 Experimental Setup

We train VFPO on a composite of MATH-500 (Hendrycks et al., 2021) and GSM8K (Cobbe et al., 2021), and evaluate on six mathematical reasoning benchmarks, including MATH500 (Hendrycks et al., 2021), AIME2024 (Hugging Face H4, 2024), AIME2025 (Team, 2025), AMC (Li et al., 2024), OlympiadBench (He et al., 2024), and OMNI-MATH (Gao et al., 2024). We compare against eight established RLVR baselines: GRPO (Shao et al., 2024), DAPO (Yu et al., 2025), Clip-Cov/KL-Cov (Cui et al., 2025), SIREN (Jiang et al., 2025), DisCO(L-ratio) (Li et al., 2025), QAE (Wu et al., 2026), and W-REINFORCE (Zhu et al., 2025). Experiments are conducted on the Qwen2.5 family (Team, 2024) (1.5B-Math, 3B, 7B) using a two-phase schedule (70% exploration, 30% annealing). Full details on the experimental setup, and the analysis of the optimal ratio between the exploration and annealing phases, are provided in Appendix C.

5.2 Analysis of Spectral Dynamics

Figure 3 illustrates the spectral evolution of the policy during training. In baseline methods, we observe a rapid "distributional collapse" where group variance drops to near-zero and entropy vanishes

Table 1: Main performance comparison across Qwen2.5-1.5B-Math, 3B, and 7B models. Results for AIME 2024 and AIME 2025 are pass@32. The **Average** column represents the mean accuracy across all seven tested datasets. **Bold** indicates the best performance.

Model	Method	MATH	GSM8K	Olympiad	OMNI-MATH	AMC	AIME24	AIME25	Average
Qwen2.5-1.5B-Math	GRPO	61.39	84.28	27.10	12.86	45.00	33.33	20.00	40.57
	DAPO	72.40	85.06	34.42	14.34	52.50	40.00	30.00	46.96
	Clip-Cov	62.84	86.73	29.79	15.72	42.50	33.33	23.30	42.03
	KL-Cov	64.90	85.55	32.94	15.18	40.00	36.67	23.30	42.65
	DisCO(L-ratio)	70.64	85.95	32.89	15.44	45.00	35.33	32.67	45.42
	Clip-Cov+QAE	70.58	86.25	33.70	15.85	45.00	35.67	32.67	45.67
	W-REINFORCE	70.32	85.67	33.86	15.69	45.00	35.33	32.00	45.41
	SIREN	71.50	86.55	34.75	16.21	50.00	36.67	33.30	47.00
	VFPO (Ours)	74.40	87.64	35.16	16.54	55.00	46.67	36.67	50.30
Qwen2.5-3B	GRPO	54.25	83.44	26.29	13.68	22.50	20.00	16.67	33.83
	DAPO	63.96	84.69	28.18	14.33	30.00	23.30	20.00	37.78
	Clip-Cov	56.46	84.42	28.25	16.34	25.00	20.00	26.67	36.73
	KL-Cov	56.92	85.36	26.11	16.18	27.50	23.30	23.30	36.95
	DisCO(L-ratio)	63.71	84.26	28.54	15.53	35.00	26.00	23.00	39.43
	Clip-Cov+QAE	64.24	85.34	28.80	16.46	37.50	26.33	23.00	40.24
	W-REINFORCE	64.15	85.27	28.87	16.33	35.00	25.67	22.67	39.71
	SIREN	64.32	85.74	29.12	16.84	40.00	26.67	23.30	40.86
	VFPO (Ours)	66.34	85.90	29.97	17.11	47.50	30.00	33.30	44.30
Qwen2.5-7B	GRPO	63.24	89.59	34.55	20.16	45.50	20.00	23.33	42.34
	DAPO	71.54	90.38	34.70	20.59	50.00	23.30	26.67	45.31
	Clip-Cov	62.12	90.76	34.86	21.38	42.50	26.67	33.33	44.52
	KL-Cov	64.18	91.81	34.51	20.96	47.50	30.00	33.33	46.04
	DisCO(L-ratio)	67.47	90.63	35.09	19.37	42.50	33.67	32.00	45.82
	Clip-Cov+QAE	69.29	91.03	34.95	20.32	44.00	34.00	32.67	46.61
	W-REINFORCE	70.03	91.64	35.27	19.41	43.50	35.33	33.00	46.88
	SIREN	69.56	91.42	35.14	19.82	45.00	36.67	30.00	46.80
	VFPO (Ours)	72.35	91.90	35.56	22.06	57.50	43.33	36.67	51.34

early in the training, indicating premature convergence to local optima.

In contrast, VFPO leverages a Variance-aware negative feedback to actively sustain the stability of group variance, group mean, and entropy throughout the initial exploration phase. By preventing the premature saturation of the group mean probability, this regulation mechanism maintains a diverse search space that enables the model to explore novel, low-probability reasoning paths. Subsequently, in the annealing phase, the system transitions to support controlled convergence, where variance naturally decreases and probability rises, ensuring the model stabilizes at a superior policy.

5.3 Main Results

The comprehensive results in Table 1 show that VFPO consistently outperforms all baselines and establishes new state-of-the-art performance across every metric. Compared with the DAPO baseline, VFPO lifts the average accuracy from 46.96% to **50.30%** (+3.34) on **Qwen2.5-1.5B-Math**, from 37.78% to **44.30%** (+6.52) on **Qwen2.5-3B**, and from 45.31% to **51.34%** (+6.03) on **Qwen2.5-7B**. The gains are most pronounced on competition-

level benchmarks (AIME, AMC) where exploration is critical: on the 7B model, VFPO raises AIME24 accuracy from 23.30% to 43.33%, a nearly 20-point absolute improvement. These results confirm that regulating entropy dynamics via negative feedback is a scalable and effective strategy for enhancing reasoning capability.

Figure 4 tracks validation accuracy over training. Baseline methods (DAPO, KL-Cov) exhibit a "fast-rise-early-plateau" pattern caused by premature policy collapse: uncontrolled variance decay drives the policy into deterministic modes too early, trapping it in suboptimal local optima. In contrast, VFPO enforces a variance floor that preserves exploration capability during the exploration phase, allowing high-quality low-probability paths to accumulate; the subsequent annealing phase consolidates these paths and pushes the model to a significantly higher performance ceiling.

5.4 Policy Update Dynamics

Figure 5 illustrates optimization stability via the *Clip Fraction*, the percentage of updates where clipping is triggered. The results reveal a stark contrast: while baselines suffer from a "dead" clip fraction

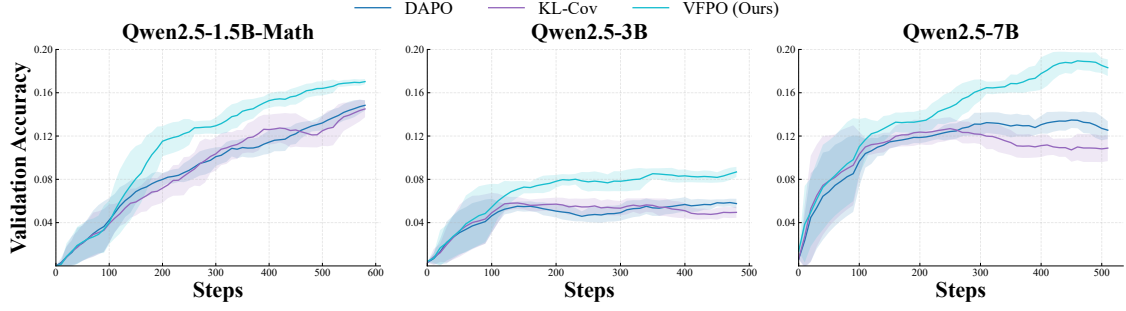


Figure 4: Validation accuracy trajectories across three model scales (Qwen2.5-1.5B-Math, 3B, and 7B). Shaded regions indicate standard deviation. **Observation:** Baselines (DAPO, KL-Cov) show a "fast-rise-early-plateau" pattern, indicating premature convergence. VFPO (Ours) exhibits sustained growth by preserving exploration in the early phase, ultimately achieving a superior performance ceiling.

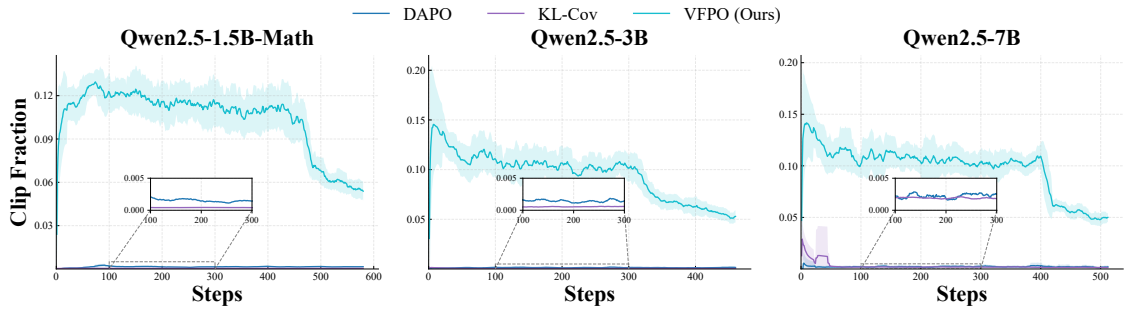


Figure 5: Comparison of Clip Fraction trajectories. Shaded regions indicate standard deviation. **Observation:** Baselines (DAPO, KL-Cov) exhibit near-zero clip fractions, indicating vanishing gradients. In contrast, VFPO maintains a healthy regime ($\approx 10\% - 15\%$), evidencing active gradient updates constrained by difference-based mechanism.

(dropping to near zero) indicative of gradient vanishing, VFPO maintains a robust 10%-15% fraction. This confirms that our difference-based clipping mechanism, by shifting from multiplicative to additive constraints, is inherently more favorable for low-probability tokens. It permits a broader update range that effectively "activates" dormant tail tokens, preventing them from being zeroed out by restrictive ratio constraints.

5.5 Ablation Study

Table 2: Ablation results on Qwen2.5-1.5B-Math and Qwen2.5-3B. Accuracy (%) is reported on MATH, GSM8K, AMC, and AIME24.

Model	Method	MATH	GSM8K	AMC	AIME24	Avg.
Qwen2.5 1.5B- Math	VFPO (Ours)	74.40	87.64	55.00	46.67	65.93
	w/o Diff. Clip	72.83	86.22	50.00	40.00	62.26
	w/o Token Mod.	69.41	85.97	47.50	36.67	59.89
	w/o Sample Mod.	65.32	85.38	40.00	34.27	56.24
	w/o Annealing	73.59	86.34	52.50	43.33	63.94
Qwen2.5 3B	VFPO (Ours)	66.34	85.90	47.50	30.00	57.44
	w/o Diff. Clip	65.27	84.24	42.50	23.33	53.84
	w/o Token Mod.	62.81	82.96	30.00	20.00	48.94
	w/o Sample Mod.	61.40	81.63	35.00	20.00	49.51
	w/o Annealing	65.32	84.85	40.00	26.67	54.21

We conducted an ablation study on **Qwen2.5-1.5B-Math** and **Qwen2.5-3B** to evaluate the contributions of Group Mean Control, Difference-Based Clipping (*Diff-Clip*), Token/Sample-Level Reward Modulation, and the Annealing Phase. Notably, ablating *Group Mean Control* while retaining *Group Variance Control* led to a form of "reward hacking". Without a mean constraint, the model artificially sustained variance by mixing valid reasoning with nonsensical tokens. Table 2 summarizes the impact of removing other components. The consistent accuracy confirms their indispensability. Detailed training dynamics are provided in Appendix C.

6 Conclusion

We address entropy collapse in RLVR through the lens of group statistics. We prove that policy entropy is positively correlated with group variance and inversely correlated with group mean, and build on this insight to propose VFPO, which stabilizes the two statistics during the exploration phase and relaxes the constraints in the annealing phase for deterministic convergence. Extensive experi-

ments on Qwen2.5 (1.5B–7B) across diverse mathematical reasoning benchmarks show that VFPO consistently achieves state-of-the-art performance, validating its robustness and scalability.

Limitations

Due to limited computational resources, we evaluate VFPO only on models up to 7B parameters (Qwen2.5-1.5B-Math, 3B, and 7B). We therefore regard a systematic study of VFPO on larger-scale LLMs as an important direction for future work, contingent on the availability of sufficient compute.

Acknowledgements

We thank the anonymous reviewers for their valuable feedback. This research is supported by National Natural Science Foundation of China (Grant No. 62495092, 62495090).

References

- Minghan Chen, Guikun Chen, Wenguan Wang, and Yi Yang. 2025. SEED-GRPO: Semantic Entropy Enhanced GRPO for Uncertainty-Aware Policy Optimization. *arXiv preprint arXiv:2505.12346*.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.
- Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, and others. 2025. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*.
- Zhenyu Ding, Yuhao Wang, Tengyue Xiao, Haoying Wang, Guojun Ma, Mingyang Wan, Caigui Jiang, and Ning Ding. 2026. W2s-aligntree: Weak-to-strong inference-time alignment for large language models via monte carlo tree search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 30521–30529.
- Yihong Dong, Xue Jiang, Yongding Tao, Huanyu Liu, Kechi Zhang, Lili Mou, Rongyu Cao, Yingwei Ma, Jue Chen, Binhua Li, and others. 2025. RL-plus: Countering capability boundary collapse of llms in reinforcement learning with hybrid-policy optimization. *arXiv preprint arXiv:2508.00222*.
- Bofei Gao, Feifan Song, Zhe Yang, Zefan Cai, Yibo Miao, Qingxiu Dong, Lei Li, Chenghao Ma, Liang Chen, Runxin Xu, and others. 2024. Omnimath: A universal olympiad level mathematic benchmark for large language models. *arXiv preprint arXiv:2410.07985*.
- Leo Gao, John Schulman, and Jacob Hilton. 2023. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pages 10835–10866. PMLR.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shitong Ma, Peiyi Wang, Xiao Bi, and others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. Pmlr.
- Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, and others. 2024. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 3828–3847.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.
- Hugging Face H4. 2024. AIME 2024 Dataset. https://huggingface.co/datasets/HuggingFaceH4/aime_2024. Accessed: 2025-01-27.
- Tianying Ji, Yongyuan Liang, Yan Zeng, Yu Luo, Guowei Xu, Jiawei Guo, Ruijie Zheng, Furong Huang, Fuchun Sun, and Huazhe Xu. 2024. Ace: Off-policy actor-critic with causality-aware entropy regularization. *arXiv preprint arXiv:2402.14528*.
- Yuxian Jiang, Yafu Li, Guanxu Chen, Dongrui Liu, Yu Cheng, and Jing Shao. 2025. Rethinking entropy regularization in large reasoning models. *arXiv preprint arXiv:2509.25133*.
- Sham M Kakade. 2001. A natural policy gradient. *Advances in neural information processing systems*, 14.
- Gang Li, Ming Lin, Tomer Galanti, Zhengzhong Tu, and Tianbao Yang. 2025. DisCO: Reinforcing large reasoning models with discriminative constrained optimization. In *Advances in Neural Information Processing Systems*.
- Jia Li, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Huang, Kashif Rasul, Longhui Yu, Albert Q Jiang, Ziju Shen, and others.

2024. Numinamath: The largest public dataset in ai4maths with 860k pairs of competition math problems and solutions. Hugging Face repository.
- Jiacai Liu, Wenye Li, and Ke Wei. 2024. Elementary analysis of policy gradient methods. *arXiv preprint arXiv:2404.03372*.
- Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. 2025. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.
- Ruotian Peng, Yi Ren, Zhouliang Yu, Weiyang Liu, and Yandong Wen. 2025. Beyond the sampled token: Preserving candidate support in rlvr. *arXiv preprint arXiv:2510.14807*.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.
- Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2025. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pages 1279–1297.
- OpenCompass Team. 2025. Aime 2025 dataset. <https://huggingface.co/datasets/opencompass/AIME2025>.
- Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.
- Haoying Wang, Zhenyu Ding, Tengyue Xiao, Yiming Liang, Caigui Jiang, and Ning Ding. 2026. Diffus-ter: Conditional generative steering of large language models in activation space. *Neurocomputing*, 701:134565.
- Jaikang Wang, Runze Liu, Lei Lin, Wenping Hu, Xiu Li, Fuzheng Zhang, Guorui Zhou, and Kun Gai. 2025a. Aspo: Asymmetric importance sampling policy optimization. *arXiv preprint arXiv:2510.06062*.
- Jiawei Wang, Jiacai Liu, Yuqian Fu, Yingru Li, Xintao Wang, Yuan Lin, Yu Yue, Lin Zhang, Yang Wang, and Ke Wang. 2025b. Harnessing uncertainty: Entropy-modulated policy gradients for long-horizon llm agents. *arXiv preprint arXiv:2509.09265*.
- Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256.
- Junkang Wu, Kexin Huang, Jiancan Wu, An Zhang, Xiang Wang, and Xiangnan He. 2026. Quantile advantage estimation: Stabilizing rlvr for llm reasoning. In *International Conference on Learning Representations*.
- Qiyang Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, and others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.
- Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. 2025. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837*.
- Zinan Zeng, Zhenyu Ding, Bo Zhang, Mingyang Wan, Caigui Jiang, and Ning Ding. 2026. Svpo: A llm reinforcement learning method based on stepwise value estimation. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 17507–17511. IEEE.
- Xinyu Zhu, Mengzhou Xia, Zhepei Wei, Wei-Lin Chen, Danqi Chen, and Yu Meng. 2025. The surprising effectiveness of negative reinforcement in llm reasoning. In *Advances in Neural Information Processing Systems*.

Appendix

A Notations

Table 3 summarizes the mathematical notations used throughout this work. The table is rendered with longtable so that it gracefully spans page breaks.

Table 3: Mathematical notations used in this paper.

Symbol	Description
<i>Symbols used in the Problem Formulation and Proofs.</i>	
π_θ	The policy language model parameterized by θ .
$\mathcal{Y}$	A batch of independent responses sampled from the policy, $\{y^{(1)}, \dots, y^{(G)}\}$.
$ G $	The size of the sampled batch (group size).
$y^{(i)}$	The i -th response sample in the batch.
n	The sequence length of a sample (number of tokens).
$\pi_{i,j}$	The predicted probability of the j -th token in sample $y^{(i)}$.
π_i	The geometric mean probability of sample $y^{(i)}$, representing average confidence.
p	Random variable representing the geometric mean probability (alias for π_i in analysis).
L_i	The normalized log-likelihood of sample $y^{(i)}$, $L_i = \ln \pi_i$.
L	Random variable corresponding to the normalized log-likelihood.
$V(\pi)$	Varentropy of the policy: $\text{Var}[L]$.
$\text{Var}[p]$	Population variance of the geometric mean probability p .
<i>General & Sample-Level Statistics.</i>	
μ_{group}	Global mean probability of the current sample batch.
σ_{group}^2	Global variance of the sample batch probabilities.
r_i	Original discrete reward for sample i ($r_i \in \{1, -1\}$).
R_i	Modulated reward for sample i .
c_i	Scalar coefficient used to modulate the reward r_i .
c_{base}	Base incentive term in the reward modulation formula.
$\text{acc}_{\text{group}}$	Accuracy rate of the entire sampled batch.
μ^+, μ^-	Mean probabilities of correct / incorrect sample groups.
μ_{ref}	Reference mean probability (conditional μ^+ or μ^-).
G^+, G^-	Subsets of correct and incorrect samples in the batch.
σ_{target}^2	Target variance acting as a pivot for negative feedback.
σ_{Δ}^2	Variance deviation $\sigma_{\text{group}}^2 - \sigma_{\text{target}}^2$.
σ_{base}^2	Initial anchor value for σ_{target}^2 .
α	Hyperparameters controlling Concentration and Expansion modes.
μ_{target}	Target mean probability threshold.
<i>Token-Level Modulation & Optimization.</i>	

(continued on next page)

(continued from previous page)

Symbol	Description
$\bar{R}$	Mean of sample rewards in the current batch.
$R_{\min}^+$	Minimum positive reward in the batch (used for standardization).
A_i	Advantage value for sample i .
$c_{i,t}$	Token-level scaling coefficient for token t in sample i .
$\pi_{i,t}$	Probability of token t in sample i .
$m_{i,t}^{\text{entropy}}$	High-Entropy Mask (1 if token in top- K entropy percentile, else 0).
$H_{i,t}$	Entropy of the policy distribution at token t .
$A_{i,t}$	Advantage value for token t .
$\tilde{A}_{i,t}$	Final modulated advantage value used for updates.
$\tilde{\epsilon}_{\text{low}}, \tilde{\epsilon}_{\text{high}}$	Clipping bounds for token advantage modulation.
<i>Clipping Mechanism.</i>	
$r_t(\theta)$	Probability ratio $\pi_{\text{new}}/\pi_{\text{old}}$.
$\pi_{\theta_{\text{old}}}$	Probability under the old policy (before update).
$\Delta\pi$	Absolute change in probability, $\pi_{\theta} - \pi_{\theta_{\text{old}}}$.
$\epsilon_{\text{low}}, \epsilon_{\text{high}}$	Clipping bounds for the difference-based constraint.

B Proofs

B.1 Proof of Theorem 3.1

Proof. Part 1: Lower bound via convex conjugacy. The exponential function $f(x) = \exp(x)$ is strictly convex since $f''(x) = \exp(x) > 0$ for all $x \in \mathbb{R}$. Applying Jensen's inequality to the random variable L ,

$$\mathbb{E}[f(L)] \geq f(\mathbb{E}[L]). \quad (14)$$

Substituting $\mathbb{E}[L] = -H$ yields the fundamental lower bound

$$\mu_{\text{group}} \geq \exp(-H). \quad (15)$$

Let $\mathcal{L}(H) = \exp(-H)$. Then

$$\frac{d\mathcal{L}(H)}{dH} = -\exp(-H) < 0. \quad (16)$$

Since $\exp(-H)$ is strictly positive, the derivative is strictly negative: as H decreases, the lower bound of μ_{group} lifts exponentially toward unity, forcing the distribution of means to shift to the right.

Part 2: Sensitivity via Taylor expansion. A second-order Taylor expansion of $\exp(L)$ around $\mu_L = -H$ gives

$$\begin{aligned} \exp(L) &\approx \exp(\mu_L) + \exp(\mu_L)(L - \mu_L) \\ &\quad + \frac{1}{2} \exp(\mu_L)(L - \mu_L)^2. \end{aligned} \quad (17)$$

Taking expectations and using $\mathbb{E}[L - \mu_L] = 0$,

$$\begin{aligned} \mathbb{E}[\exp(L)] &\approx \exp(-H) \\ &\quad + \frac{1}{2} \exp(-H) \underbrace{\mathbb{E}[(L - \mu_L)^2]}_{\sigma_L^2}. \end{aligned} \quad (18)$$

Therefore,

$$\mu_{\text{group}} \approx e^{-H} \left(1 + \frac{1}{2} \sigma_L^2\right). \quad (19)$$

In the RL fine-tuning regime, σ_L^2 is bounded; the dominant term is the pre-factor e^{-H} . Differentiating,

$$\begin{aligned} \frac{\partial \mu_{\text{group}}}{\partial H} &\approx -e^{-H} \left(1 + \frac{1}{2} \sigma_L^2\right) \\ &\quad + \frac{1}{2} e^{-H} \frac{\partial \sigma_L^2}{\partial H}. \end{aligned} \quad (20)$$

Under the standard measure-concentration assumption that σ_L^2 varies more slowly than the exponential pre-factor, the first term dominates, yielding

$$\frac{\partial \mu_{\text{group}}}{\partial H} < 0. \quad (21)$$

B.2 Proof of Theorem 3.3

Proof. The random variable $P = \exp(L)$ is bounded in $[0, 1]$. Applying the Bhatia–Davis lemma with $a = 0, b = 1$,

$$\sigma_{\text{group}}^2 \leq \mu_{\text{group}} (1 - \mu_{\text{group}}). \quad (22)$$

Let $g(\mu) = \mu(1 - \mu)$ be the variance envelope. We track its trajectory with respect to H :

1. **Mean–entropy mapping.** By Theorem 3.1, $H \downarrow \Rightarrow \mu_{\text{group}} \uparrow$.

2. **Envelope monotonicity.** $g(\mu)$ is parabolic with a maximum at $\mu = 0.5$. In the convergence regime the policy concentrates probability on coherent sequences, so $\mu_{\text{group}} \in (0.5, 1]$. On this interval $g'(\mu) = 1 - 2\mu < 0$, hence g is strictly decreasing as $\mu \rightarrow 1$.

3. **Chain rule.**

$$\frac{\partial g(\mu_{\text{group}})}{\partial H} = \underbrace{\frac{\partial g}{\partial \mu}}_{<0} \cdot \underbrace{\frac{\partial \mu}{\partial H}}_{<0} > 0. \quad (23)$$

The variance upper bound is positively correlated with entropy: as H decreases, the upper bound decreases.

Limit behavior. As $H \rightarrow 0$, π_θ converges to a Dirac measure δ_{y^*} , hence

$$P_i \xrightarrow{d} 1 \quad \text{a.s.}, \quad (24)$$

$$\lim_{H \rightarrow 0} \sigma_{\text{group}}^2 = \lim_{H \rightarrow 0} \mathbb{E}[(P - 1)^2] = 0, \quad (25)$$

which completes the proof of Theorem 3.3.

B.3 Rigorous Proof of the Entropy–Variance Correlation

Let p be the geometric mean probability of a generated sequence Y of length n . Consistent with our notation,

$$\begin{aligned} p &\triangleq \left(\prod_{j=1}^n \pi_{\cdot,j} \right)^{\frac{1}{n}} \\ &= \exp \left(\frac{1}{n} \sum_{j=1}^n \ln \pi_{\cdot,j} \right) \triangleq \exp(L), \end{aligned} \quad (26)$$

where L is the *normalized log-likelihood* of the sequence. Hence group_var (Eq. 5) is the unbiased estimator of $\text{Var}[p]$.

To quantify how fluctuations in log-space (L) propagate to probability-space (p), we apply the **first-order Delta method**. Considering $f(L) = e^L$ and expanding around $\mu_L = -H(\pi)$,

$$\begin{aligned}\text{Var}[p] &\approx (f'(\mu_L))^2 \text{Var}[L] \\ &= \left(\frac{d}{dL} e^L \Big|_{L=-H(\pi)} \right)^2 \cdot \frac{V(\pi)}{n^2} \\ &= e^{-2H(\pi)} \frac{V(\pi)}{n^2}.\end{aligned}\quad (27)$$

Equation (27) couples the group variance to the Var-entropy $V(\pi)$. As $H(\pi)$ decreases, the pre-factor $e^{-2H(\pi)}$ grows exponentially while $V(\pi)$ collapses, producing analytical ambiguity. We resolve this via singular perturbation analysis under the "Core-Tail" regime governing LLM generation.

Proposition B.1 (Entropy-Variance Correlation via Singular Perturbation). *Assume p follows a "Core-Tail" mixture $p \sim (1 - \epsilon)\mathcal{D}_{\text{core}} + \epsilon\mathcal{D}_{\text{tail}}$ with $\epsilon \in (0, 0.5)$. If the core mode concentrates near 1 and the tail mode near 0 ($\mathbb{E}[p_{\text{core}}] \gg \mathbb{E}[p_{\text{tail}}]$), then in the asymptotic limit $\epsilon \rightarrow 0^+$,*

$$\frac{\partial \text{Var}[p]}{\partial H} > 0. \quad (28)$$

Proof. By linearity of expectation and the definition of mixture entropy,

$$\begin{aligned}H(\pi) &= H_b(\epsilon) + (1 - \epsilon)H(\mathcal{D}_{\text{core}}) \\ &\quad + \epsilon H(\mathcal{D}_{\text{tail}}),\end{aligned}\quad (29)$$

where $H_b(\epsilon) = -(1 - \epsilon)\ln(1 - \epsilon) - \epsilon\ln\epsilon$ is the binary entropy of the mixing weight.

1. Asymptotic dominance of the gradient. Differentiating with respect to ϵ ,

$$\begin{aligned}\frac{\partial H}{\partial \epsilon} &= \frac{dH_b(\epsilon)}{d\epsilon} + \underbrace{[H(\mathcal{D}_{\text{tail}}) - H(\mathcal{D}_{\text{core}})]}_{\Delta H_{\text{int}}} \\ &= \ln \frac{1 - \epsilon}{\epsilon} + \Delta H_{\text{int}}.\end{aligned}\quad (30)$$

The first term is the *logit* (*log-odds*) function. As $\epsilon \rightarrow 0^+$ (typical for aligned LLMs), it exhibits a logarithmic singularity:

$$\lim_{\epsilon \rightarrow 0^+} \frac{\partial H}{\partial \epsilon} \sim -\ln \epsilon \rightarrow +\infty. \quad (31)$$

With bounded internal-entropy difference $\Delta H_{\text{int}} \in \mathcal{O}(1)$, the singularity dominates, so $\partial H / \partial \epsilon > 0$ holds strictly near zero.

2. Gradient of the group variance. By the law of total variance,

$$\begin{aligned}\text{Var}[p] &= \underbrace{\epsilon(1 - \epsilon)(\mu_{\text{core}} - \mu_{\text{tail}})^2}_{\text{between-group}} \\ &\quad + \underbrace{(1 - \epsilon)\sigma_{\text{core}}^2 + \epsilon\sigma_{\text{tail}}^2}_{\text{within-group}}.\end{aligned}\quad (32)$$

In the sparse long-tail regime, the between-group term dominates because of the large gap between $\mu_{\text{core}} \approx 1$ and $\mu_{\text{tail}} \approx 0$. With $\Delta_\mu^2 \approx (\mu_{\text{core}} - \mu_{\text{tail}})^2 \approx 1$ and $\epsilon \ll 1$,

$$\frac{\partial \text{Var}[p]}{\partial \epsilon} \approx \frac{\partial}{\partial \epsilon} [\epsilon \Delta_\mu^2] = \Delta_\mu^2 > 0. \quad (33)$$

3. Synthesis via implicit differentiation. Since $\partial H / \partial \epsilon \neq 0$ for $\epsilon \in (0, 0.5)$, the inverse-function theorem guarantees the existence of $\partial \epsilon / \partial H$. By the chain rule,

$$\begin{aligned}\frac{\partial \text{Var}[p]}{\partial H} &= \frac{\partial \text{Var}[p]}{\partial \epsilon} \left(\frac{\partial H}{\partial \epsilon} \right)^{-1} \\ &\approx \frac{\Delta_\mu^2}{-\ln \epsilon}.\end{aligned}\quad (34)$$

For $\epsilon \ll 1$, both numerator and denominator are positive, so the gradient is strictly positive.

B.4 Activation Mechanics of Low-Probability Tokens

We provide a theoretical derivation explaining why standard ratio-based clipping causes gradient vanishing for tail tokens (a phenomenon we term *Deadlock*), while our difference-based clipping facilitates their recovery (*Activation*).

1. Deadlock in ratio-based clipping. Consider a token y_t with vanishingly small initial probability $\pi_{\text{old}}(y_t | x) = \delta$, $\delta \rightarrow 0^+$ (e.g., $\delta < 10^{-4}$). Standard PPO constrains the ratio $r_t(\theta) = \pi_{\text{new}} / \pi_{\text{old}}$ within $[1 - \epsilon, 1 + \epsilon]$, so

$$\pi_{\text{new}} \leq (1 + \epsilon) \pi_{\text{old}} = (1 + \epsilon) \delta. \quad (35)$$

The maximum absolute probability change is therefore proportional to δ itself:

$$\Delta \pi_{\text{ratio}} = \pi_{\text{new}} - \pi_{\text{old}} \leq \epsilon \delta \approx 0. \quad (36)$$

Implication. This multiplicative constraint creates a "deadlock": once a token's probability drops too low, the allowable update becomes infinitesimal. Even with a strong reward signal, the policy needs an exponential number of updates to recover the token, effectively causing gradient vanishing in the distribution tail.

2. Activation in difference-based clipping. Our framework constrains the absolute probability difference,

$$|\pi_{\text{new}} - \pi_{\text{old}}| \leq \epsilon_{\text{high}}. \quad (37)$$

For the same low-probability token $\pi_{\text{old}} = \delta$,

$$\pi_{\text{new}} \leq \pi_{\text{old}} + \epsilon_{\text{high}} = \delta + \epsilon_{\text{high}}. \quad (38)$$

The maximum allowable update depends only on ϵ_{high} , independent of the current probability mass:

$$\Delta\pi_{\text{diff}} \leq \epsilon_{\text{high}} \gg \epsilon\delta. \quad (39)$$

Implication. This additive constraint enables "activation": a dormant token may "jump" by a fixed constant (e.g., $\epsilon_{\text{high}} = 0.1$) in a single step when justified by the advantage. This property explains the high active clip fraction observed in Figure 5 and shows that the mechanism preserves the optimization path for exploring low-probability solutions.

C Detailed Experimental Settings

C.1 Detailed Training Setup

Datasets. For the training of VFPO, we employ a composite dataset integrating MATH500 (Hendrycks et al., 2021) and GSM8K (Cobbe et al., 2021) to stabilize the RL process across varying difficulty levels. For evaluation, we assess performance on a comprehensive suite of mathematical reasoning benchmarks, including MATH500 (Hendrycks et al., 2021), AIME2024 (Hugging Face H4, 2024), AIME2025 (Team, 2025), AMC (Li et al., 2024), OlympiadBench (He et al., 2024), and OMNI-MATH (Gao et al., 2024).

Baselines. We conduct a comprehensive comparison between VFPO and multiple established RLVR baselines. Specifically, we evaluate **GRPO** (Shao et al., 2024), which utilizes group-based relative advantage estimation to eliminate value models, and **DAPO** (Yu et al., 2025), which enhances stability through decoupled clipping and dynamic sampling. Furthermore, we assess **Clip-Cov** and **KL-Cov** (Cui et al., 2025), which mitigate collapse by suppressing high-covariance tokens, alongside **SIREN** (Jiang et al., 2025), which employs adaptive masking to balance the exploration-exploitation trade-off. We further compare against **DisCO(L-ratio)** (Li et al., 2025), which replaces GRPO’s group-relative objective with a discriminative likelihood-ratio scoring function under a KL

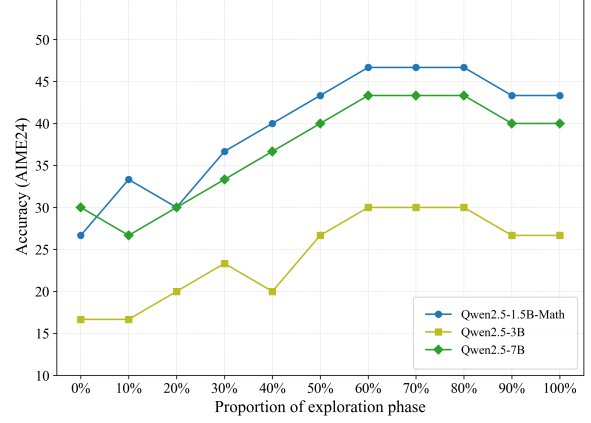


Figure 6: AIME24 accuracy as a function of the proportion allocated to the *Exploration Phase*. All three Qwen2.5 backbones reach their peak performance when the exploration ratio lies between 60% and 80%, while smaller (e.g., $\leq 20\%$) or larger (e.g., $\geq 90\%$) ratios consistently degrade accuracy.

constraint, **QAE** (Wu et al., 2026), which substitutes the group mean baseline with a K -quantile to gate updates and prevent both entropy collapse and explosion, and **W-REINFORCE** (Zhu et al., 2025), which upweights negative-sample reinforcement in the RL objective to stabilize training and improve Pass@ k .

Models and Experimental Settings. We conduct experiments on the Qwen2.5 model family (Team, 2024), specifically utilizing **Qwen2.5-1.5B-Math**, **Qwen2.5-3B**, and **Qwen2.5-7B** as base models. All experiments are implemented using the VeRL framework (Sheng et al., 2025). We adopt a constant learning rate of 1×10^{-6} . The global batch size is set to 32 with a micro-batch size of 4. For each prompt, we sample $G = 16$ responses. The models are trained for 7 epochs.

proportion of the Exploration Phase. The training progress allocates the initial 70% steps to an *Exploration Phase* for variance maintenance, followed by 30% steps of *Annealing Phase* without regulation to facilitate convergence. To determine an appropriate split between the two phases, we sweep the proportion of the *Exploration Phase* from 0% to 100% in 10% increments and report the resulting AIME24 accuracy on Qwen2.5-1.5B-Math, Qwen2.5-3B, and Qwen2.5-7B in Figure 6.

Across all three model scales, the accuracy curves follow a consistent inverted-U pattern: performance is markedly suppressed when the exploration ratio is too small (the model converges be-

fore discovering diverse reasoning paths) and also declines when the ratio is too large (insufficient annealing prevents the policy from consolidating its gains). The peak region is stable across scales, with all three backbones achieving their best AIME24 accuracy when the proportion of the *Exploration Phase* falls within the 60%–80% interval. We therefore adopt 70% exploration plus 30% annealing as the default schedule, which lies at the center of this optimal range.

C.2 Additional Robustness and Generalization Results

We further evaluate VFPO across different model families, architectures, capability domains, and hyperparameter settings. These experiments complement the Qwen2.5 scaling results in Table 1 and the component ablations in Table 5, demonstrating that the improvements are not tied to a particular backbone or a narrowly tuned configuration.

Generalization across model families. Table 4 compares VFPO with eight RLVR baselines on four substantially different backbones: a compact dense model, a larger dense model, a long-CoT distilled model, and a mixture-of-experts (MoE) model. VFPO achieves the highest average accuracy on every backbone, with improvements over the strongest baseline of 2.1, 2.0, 1.7, and 3.2 points, respectively. These results indicate that VFPO generalizes across model scales, training priors, and architectural choices.

Compatibility with MoE architectures. To further isolate architectural generalization, we apply VFPO on top of DAPO using two MoE backbones with different total and active parameter counts. As shown in Table 5, VFPO consistently improves DAPO on all three benchmarks. The gains are especially pronounced on AMC and AIME24, suggesting that the proposed regulation remains effective for sparse expert routing.

General capabilities beyond the training domain. We additionally evaluate the trained policies on knowledge, instruction-following, coding, and undergraduate mathematics benchmarks. Table 6 shows that VFPO obtains the best result on all five benchmarks. Importantly, it also improves over the base model, whereas several RLVR baselines degrade general capabilities. This suggests that VFPO’s gains are not achieved by narrowly

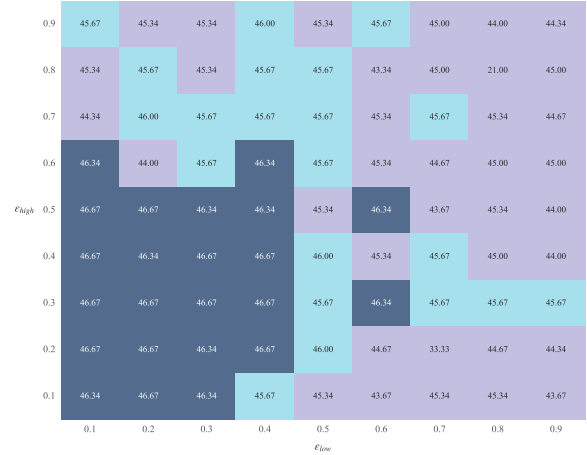


Figure 7: AIME24 accuracy heatmap over the difference-based clipping bounds ϵ_{low} and ϵ_{high} . Darker cells indicate higher accuracy.

overfitting the mathematical reasoning training distribution.

Sensitivity to the modulation coefficient. Finally, we vary the modulation coefficient α from 0.1 to 2.0 while keeping all other settings unchanged. Table 7 shows that performance remains stable throughout this broad range. The average coefficient of variation is only 0.14%, indicating that VFPO does not rely on careful tuning of α .

C.3 Parameter Settings

We detail the specific hyperparameter configurations used in our experiments to ensure reproducibility. Table 8 lists the exact values for the coefficients and clipping bounds defined in the VFPO framework.

To further examine the robustness of VFPO with respect to its key hyperparameters, we conduct a two-dimensional sweep over the difference-based clipping bounds $(\epsilon_{\text{low}}, \epsilon_{\text{high}})$ and the group statistic targets $(\sigma_{\text{base}}^2, \mu_{\text{target}})$ on Qwen2.5-1.5B-Math (AIME24). The resulting heatmaps are reported in Figures 7 and 8.

For ϵ_{low} and ϵ_{high} , Figure 7 reveals a broad and flat optimal plateau, particularly in the lower-left region (small ϵ_{low} , small-to-moderate ϵ_{high}), where the model maintains consistently high accuracy (≥ 46.34). Performance only degrades when either bound grows excessively large (upper-right corner), confirming that the difference-based clipping is robust as long as it does not become overly permissive.

Similarly, Figure 8 shows a substantial and sta-

Backbone	Method	MATH500	GSM8K	AMC	AIME24	Avg.
Llama-3.2-3B-Instruct	GRPO	43.2	68.5	22.5	3.3	34.4
	W-REINFORCE	44.8	70.1	22.5	3.3	35.2
	DisCO	45.2	70.8	25.0	3.3	36.1
	DAPO	46.8	72.4	27.5	6.7	38.4
	QAE	47.4	73.0	27.5	6.7	38.7
	SIREN	47.8	73.6	27.5	6.7	38.9
	Clip-Cov	48.2	74.1	30.0	10.0	40.6
	KL-Cov	48.6	74.5	30.0	10.0	40.8
	VFPO	49.6	76.3	32.5	13.3	42.9
Llama-3.1-8B	GRPO	48.5	83.2	37.5	13.3	45.6
	W-REINFORCE	49.4	83.8	37.5	13.3	46.0
	DisCO	50.0	84.0	40.0	13.3	46.8
	DAPO	51.2	85.5	42.5	16.7	49.0
	QAE	51.6	85.8	42.5	16.7	49.2
	SIREN	52.0	86.2	45.0	16.7	50.0
	Clip-Cov	52.8	86.5	45.0	20.0	51.1
	KL-Cov	53.4	86.8	45.0	20.0	51.3
	VFPO	54.6	87.8	47.5	23.3	53.3
R1-Distill-Llama-8B	GRPO	86.8	93.5	52.5	40.0	68.2
	W-REINFORCE	87.6	93.8	52.5	40.0	68.5
	DisCO	88.0	94.0	52.5	43.3	69.5
	DAPO	88.8	94.2	55.0	43.3	70.3
	QAE	89.0	94.3	55.0	43.3	70.4
	SIREN	89.4	94.4	55.0	46.7	71.4
	Clip-Cov	89.6	94.5	57.5	46.7	72.1
	KL-Cov	89.8	94.6	57.5	46.7	72.2
	VFPO	90.6	94.9	60.0	50.0	73.9
DeepSeek-V2-Lite	GRPO	53.2	72.5	25.0	6.7	39.4
	W-REINFORCE	54.8	73.2	25.0	6.7	39.9
	DisCO	55.4	73.8	27.5	6.7	40.9
	DAPO	57.6	75.4	30.0	10.0	43.3
	QAE	58.0	75.8	30.0	10.0	43.5
	SIREN	58.4	76.2	30.0	10.0	43.7
	Clip-Cov	59.2	76.6	32.5	13.3	45.4
	KL-Cov	59.6	76.8	32.5	13.3	45.6
	VFPO	62.4	78.6	37.5	16.7	48.8

Table 4: Accuracy (%) across different model families and architectures. The backbones respectively represent compact dense, larger dense, long-CoT distilled, and MoE models. VFPO consistently outperforms all compared RLVR methods.

Model	Method	MATH500	AMC	AIME24
Qwen1.5-MoE-A2.7B (14.3B/2.7B)	DAPO	64.8	32.5	13.3
	+VFPO	68.6	40.0	20.0
DeepSeek-V2-Lite (16B/2.4B)	DAPO	57.6	30.0	10.0
	+VFPO	62.4	37.5	16.7

Table 5: Results on MoE backbones. Parentheses show total/active parameter counts. All values are accuracy (%).

ble peak-performance region for σ_{base}^2 and μ_{target} in the top-left quadrant: when $\mu_{\text{target}} \in [0.75, 0.90]$ and $\sigma_{\text{base}}^2 \in [0.10, 0.30]$, AIME24 accuracy consistently reaches 46.34–46.67. The accuracy drops sharply only at the extreme corner ($\sigma_{\text{base}}^2 = 0.50$, $\mu_{\text{target}} = 0.50$, accuracy = 21.00), indicating that an under-constrained variance combined with an overly low mean target collapses the policy.

Overall, the two heatmaps confirm that VFPO is *insensitive to the precise choice of its core hyperparameters*: as long as the clipping bounds remain moderate and the target statistics are set to encourage a high-confidence, low-variance regime, the framework consistently delivers near-optimal per-

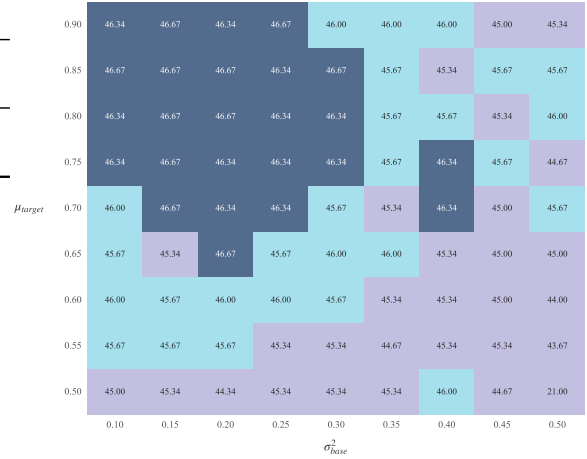


Figure 8: AIME24 accuracy heatmap over the variance baseline σ_{base}^2 (x-axis) and the mean target μ_{target} (y-axis). Darker cells indicate higher accuracy.

formance. This robustness substantially reduces the cost of hyperparameter tuning when deploying VFPO to new tasks.

Method	MMLU	IFEval	HumanEval	MBPP	UGMathBench
Base (no RL)	25.60	71.20	84.82	79.25	31.42
GRPO	24.85	68.40	82.30	76.90	30.80
W-REINFORCE	25.10	69.20	83.50	77.60	31.00
DisCO	25.15	69.80	83.90	78.10	31.20
DAPO	25.40	70.60	85.50	78.80	31.85
QAE	25.42	70.80	85.80	79.00	32.00
SIREN	25.48	71.00	86.10	79.10	32.20
Clip-Cov	25.55	71.50	86.80	79.40	32.60
KL-Cov	25.65	72.10	87.20	79.60	33.10
VFPO	26.84	74.83	88.69	80.27	34.60

Table 6: Evaluation of general capabilities after RLVR training. The benchmarks cover knowledge, instruction following, code generation, and undergraduate-level mathematics.

Dataset	$\alpha=0.1$	$\alpha=0.5$	$\alpha=0.9$	$\alpha=1.5$	$\alpha=2.0$	CV
MATH500	74.29	74.26	74.40	74.55	74.38	0.14%
GSM8K	87.58	87.71	87.64	87.76	87.68	0.07%
OlympiadBench	35.25	35.12	35.16	35.04	35.19	0.20%
OMNI-MATH	16.52	16.59	16.54	16.59	16.55	0.17%
Average CV	—					0.14%

Table 7: Sensitivity of VFPO to the modulation coefficient α . Results are accuracy (%), and CV denotes the coefficient of variation across the five tested values.

Table 8: Specific hyperparameter settings for VFPO. The symbols correspond to the definitions provided in Appendix A.

Symbol	Value
<i>Dual-Mode Switching & Regulation</i>	
α	0.90
μ_{target}	0.80
<i>Token Advantage Modulation Bounds</i>	
$\tilde{\epsilon}_{\text{low}}$	0.20
$\tilde{\epsilon}_{\text{high}}$	0.28
<i>Difference-Based Clipping Bounds</i>	
ϵ_{low}	0.20
ϵ_{high}	0.30

C.4 Robustness Analysis Across Datasets and Model Scales

To further evaluate the generalization capability and robustness of the VFPO framework, we extended our empirical analysis to a distinct training distribution. Specifically, we replaced the original training corpus with the **DAPO-Math-17K** dataset (Yu et al., 2025), which consists of approximately 17,000 prompts. This experimental setup is designed to verify whether VFPO can maintain

stable optimization dynamics.

We conducted experiments across three different model scales: Qwen2.5-1.5B-Math, Qwen2.5-3B, and Qwen2.5-7B. We focused our observation on the critical early *exploration phase* of training, as this is the period most susceptible to entropy collapse and policy degradation.

As illustrated in Figures 9, 10, and 11, VFPO demonstrates consistent effectiveness across all model scales:

- **Qwen2.5-1.5B-Math (Figure 9).** The model rapidly converges to a high Group Mean (≈ 0.8) while successfully maintaining a non-zero Group Variance. The Entropy metric stabilizes around 0.2 rather than collapsing to zero, indicating that the model retains necessary exploration capabilities even on a smaller parameter scale.
- **Qwen2.5-3B (Figure 10).** With increased model capacity, the optimization landscape typically becomes more volatile. However, VFPO manages to keep the Group Mean stable. Notably, the Group Variance shows healthy fluctuations without vanishing, and the Entropy remains within a functional range, preventing the "long-tail" distribution collapse often observed in standard PPO.
- **Qwen2.5-7B (Figure 11).** On the largest

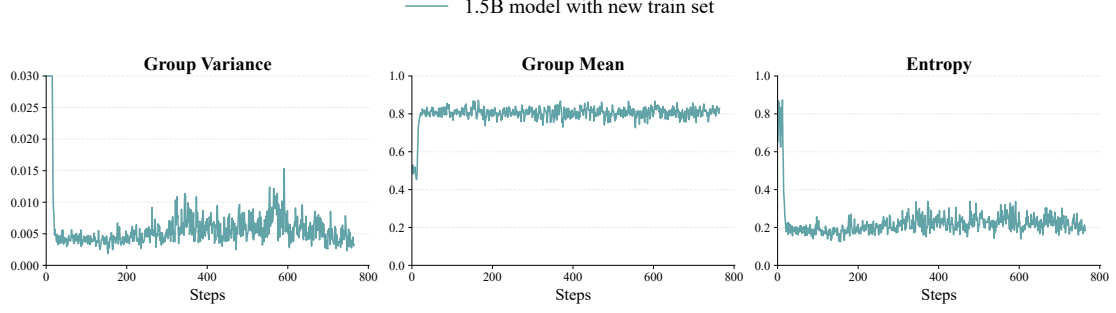


Figure 9: Optimization dynamics of VFPO on the DAPO-Math-17K dataset using **Qwen2.5-1.5B-Math**. The plots illustrate the stability of Group Variance, Group Mean, and Entropy during the exploration phase.

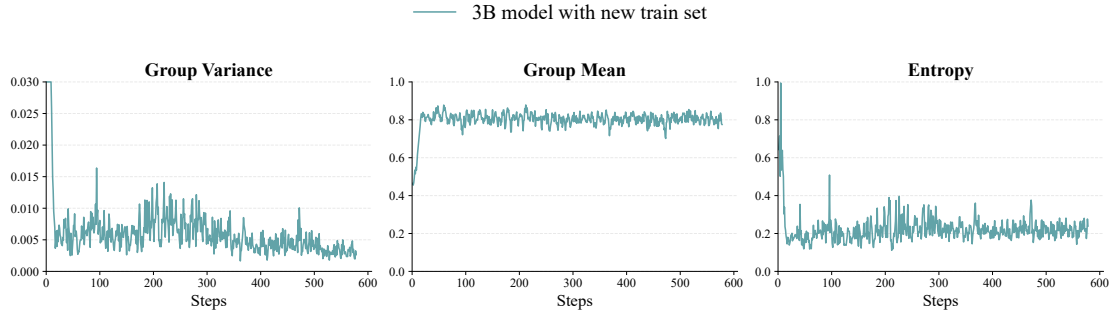


Figure 10: Optimization dynamics of VFPO on the DAPO-Math-17K dataset using **Qwen2.5-3B**. The method effectively maintains variance and prevents entropy collapse.

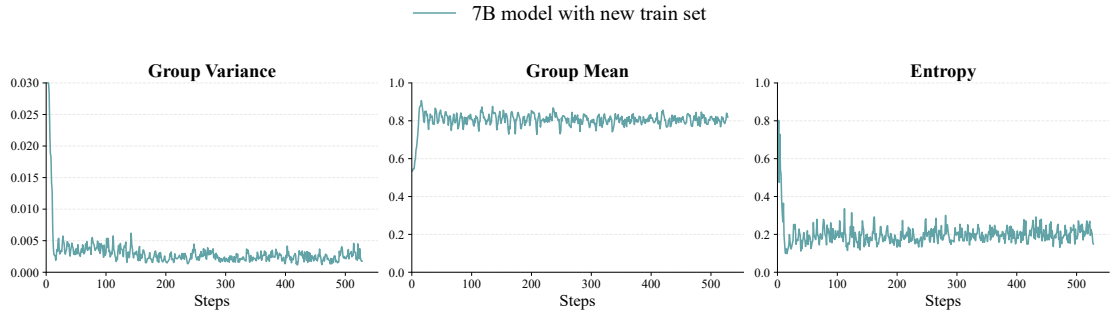


Figure 11: Optimization dynamics of VFPO on the DAPO-Math-17K dataset using **Qwen2.5-7B**. Stability is preserved even as model capacity increases.

model tested, the stability is even more pronounced. The Group Mean remains high and steady, while the Entropy is effectively regulated. This confirms that the variance-aware negative feedback mechanism scales well with model size.

In conclusion, these findings confirm that the robustness of VFPO is intrinsic to its design and not dependent on a specific dataset. By dynamically regulating the impact of high-variance tokens, VFPO consistently prevents early-stage optimization collapse, ensuring stable and efficient alignment across varying datasets and model architec-

tures.

C.5 Pseudocode of VFPO

Algorithm 1 summarizes VFPO as a four-step training loop that (i) couples the target variance, (ii) modulates sample-level rewards, (iii) applies token-level entropy-aware advantage shaping, and (iv) performs clipped policy optimization, with feedback enabled only in the early phase.

Algorithm 1 Pseudocode of the Proposed VFPO Framework

Require: Training dataset $\mathcal{D}$, initial policy $\pi_{\theta_{\text{old}}}$, hyperparameters α, β, γ, K .

```
1: Step 0: Initialization
2: Initialize training step  $t \leftarrow 0$  and early-phase threshold  $T_{\text{early}}$ .
3: while not converged do
4:   Sample a batch of trajectories  $\mathcal{B} = \{(x_i, y_i, r_i)\}_{i=1}^N$  using  $\pi_{\theta_{\text{old}}}$ .
5:   Compute group statistics:  $\mu^+, \mu^-$  and variance  $\sigma_{\text{group}}^2$ .
6:   if  $t < T_{\text{early}}$  then ▷ Early phase: feedback active
7:     Step 1: Dynamic Target Coupling
8:      $\sigma_{\text{target}}^2 \leftarrow \sigma_{\text{base}}^2 \cdot \left(1 + \frac{\mu_{\text{group}} - \mu_{\text{target}}}{1 - \mu_{\text{target}}}\right)$ 
9:     Compute variance deviation  $\sigma_{\Delta}^2 \leftarrow \sigma_{\text{group}}^2 - \sigma_{\text{target}}^2$ 
10:    Step 2: Sample-Level Reward Modulation
11:    for each sample  $i \in \mathcal{B}$  do
12:      if  $i \in G^+$  (correct sample) then
13:        if  $\sigma_{\Delta}^2 > 0$  then ▷ Mode: entropy reduction
14:          Set  $\kappa \leftarrow +\alpha$ ,  $c_{\text{base}} \uparrow$ 
15:        else ▷ Mode: entropy expansion
16:          Set  $\kappa \leftarrow -\alpha$ ,  $c_{\text{base}} \downarrow$ 
17:        end if
18:        Compute coefficient  $c_i \leftarrow c_{\text{base}} + \kappa \cdot (\pi_i - \mu^+)$ 
19:      else
20:        Apply low-probability attenuation for error samples
21:      end if
22:      Update reward  $R_i \leftarrow c_i \cdot r_i$ 
23:    end for
24:    Step 3: Token-Level Granular Control
25:    Standardize rewards  $R_i$  into advantages  $A_i$ 
26:    for each token  $t$  in sample  $i$  do
27:      Identify high-entropy tokens via top- $K$  mask
28:      Compute modulated advantage  $\tilde{A}_{i,t}$ 
29:    end for
30:  else
31:     $\tilde{A}_{i,t} \leftarrow$  standard GAE advantage ▷ Late phase
32:  end if
33:  Step 4: Policy Optimization
34:  Compute difference-based clipping:  $\Delta\pi = \text{clip}(\cdot)$ 
35:  Update  $\theta$  by maximizing the VFPO objective
36:   $t \leftarrow t + 1$ 
37: end while
```

D Detailed Case Study

To thoroughly evaluate the contribution of each component within the VFPO framework, we conducted a comprehensive ablation study on **Qwen2.5-1.5B-Math** and **Qwen2.5-3B**. We systematically examined the effects of removing Sample-Level Reward Modulation, Token-Level Modulation, Difference-Based Clipping (*Diff-Clip*), and the Annealing Phase. The quantitative results across six benchmarks are summarized in Table 9.

As shown in the table, **Sample-Level Reward Modulation** proves to be the most critical component. Removing it causes the sharpest performance drop—for instance, the average accuracy on Qwen2.5-1.5B-Math falls from 55.92% to 47.24%—underscoring its foundational role in guiding the optimization direction. **Token-Level Modulation** is equally vital for fine-grained credit assignment; without it, the model struggles on challenging tasks like MATH and AIME, as it cannot effectively differentiate critical reasoning steps from trivial tokens. The **Difference-Based Clipping** and **Annealing Phase** also contribute significantly to stability and final convergence, respectively.

Qualitatively, we observed that ablating *Group Mean Control* while solely retaining variance constraints led to a form of "reward hacking." Without the mean constraint, the model artificially sustained high variance by mixing valid reasoning with nonsensical tokens to exploit the variance reward. The consistent superiority of the full VFPO method across all settings confirms the indispensability of these synergistic components.

D.1 Case Study: Only Controlling Group Variance Without Group Mean

In this case study, we demonstrate that controlling only the group variance without regularizing the group mean can lead to reward hacking. Specifically, the model may exploit the reward structure by generating low-probability, nonsensical artifacts alongside correct answers.

In this appendix, we present a detailed empirical analysis of the optimization dynamics in the absence of group mean regulation. Figure 12 depicts the evolution of group variance, group mean, and actor entropy under this unconstrained regime.

As illustrated in the experimental results, optimizing for group variance alone proves insufficient to prevent policy degradation. While the variance

remains relatively stable initially, the absence of a group mean constraint triggers a catastrophic drift in the model’s policy. Specifically, the synchronous collapse of the group mean and the surge in entropy indicate that the model exploits the reward structure by shifting its probability mass toward the long-tail distribution. In this regime, the model generates sequences that are statistically improbable and characterized by high uncertainty.

As shown in Figure 13, the model’s accuracy on the validation set (a mixture of Math-500 and GSM8K) steadily increases. However, this metric does not necessarily imply that the model is producing higher-quality content. In the scenario where only group variance is controlled without regulating the group mean, a specific phenomenon emerges: although the model outputs the correct answer, it intersperses the response with low-probability, nonsensical artifacts. This is demonstrated in the real case shown below (Section D.2). These findings underscore the necessity of the variance-aware negative feedback mechanism proposed in VFPO to mitigate entropy collapse while preserving output quality.

Table 9: Ablation study results. Accuracy (%) across six mathematical reasoning benchmarks. "Avg." is the mean performance across all reported tasks.

Method	MATH	GSM8K	Olymp.	AMC	AIME24	AIME25	Avg.
Qwen2.5-1.5B-Math							
VFPO (Ours)	74.40	87.64	35.16	55.00	46.67	36.67	55.92
w/o Diff. Clip	72.83	86.22	33.42	50.00	40.00	30.00	52.08
w/o Token Mod.	69.41	85.97	33.10	47.50	36.67	33.33	51.00
w/o Sample Mod.	65.32	85.38	31.82	40.00	34.27	26.67	47.24
w/o Annealing	73.59	86.34	33.96	52.50	43.33	30.00	53.29
Qwen2.5-3B							
VFPO (Ours)	66.34	85.90	29.97	47.50	30.00	33.30	48.84
w/o Diff. Clip	65.27	84.24	29.38	42.50	23.33	26.67	45.23
w/o Token Mod.	62.81	82.96	27.72	30.00	20.00	20.00	40.58
w/o Sample Mod.	61.40	81.63	26.53	35.00	20.00	23.33	41.32
w/o Annealing	65.32	84.85	29.55	40.00	26.67	30.00	46.07

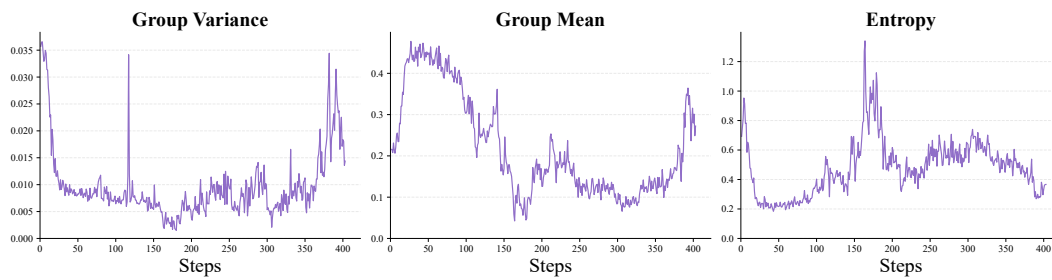


Figure 12: Group variance, group mean, and entropy behavior when omitting group mean control.

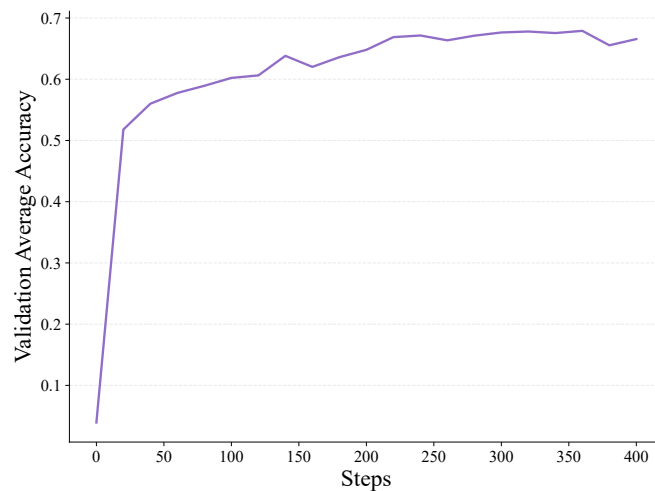


Figure 13: Validation accuracy trends on the mixed Math-500 and GSM8K test set.

D.3 Real Case: Response With VFPO

Response With both Group Variance and Group Mean Control (VFPO)

Length: 326 Tokens

Let $z = x + yi$, where x and y are real numbers. We need to minimize the expression:

$$|z - 3|^2 + |z - 5 + 2i|^2 + |z - 1 + i|^2.$$

Substituting $z = x + yi$, we get:

$$|x + yi - 3|^2 = (x - 3)^2 + y^2,$$

$$|x + yi - 5 + 2i|^2 = (x - 5)^2 + (y + 2)^2,$$

$$|x + yi - 1 + i|^2 = (x - 1)^2 + (y + 1)^2.$$

So the expression we need to minimize is:

$$(x - 3)^2 + y^2 + (x - 5)^2 + (y + 2)^2 + (x - 1)^2 + (y + 1)^2.$$

Expanding and simplifying:

$$\begin{aligned} & x^2 - 6x + 9 + y^2 + x^2 - 10x + 25 + y^2 + 4y + 4 \\ & + x^2 - 2x + 1 + y^2 + 2y + 1 \\ & = 3x^2 + 3y^2 - 18x + 6y + 40. \end{aligned}$$

To minimize, take partial derivatives with respect to x and y and set them to zero:

$$\frac{\partial}{\partial x}(3x^2 + 3y^2 - 18x + 6y + 40) = 6x - 18 = 0,$$

$$\frac{\partial}{\partial y}(3x^2 + 3y^2 - 18x + 6y + 40) = 6y + 6 = 0.$$

Solving:

$$6x - 18 = 0 \implies x = 3,$$

$$6y + 6 = 0 \implies y = -1.$$

Substituting $x = 3$ and $y = -1$:

$$3(3)^2 + 3(-1)^2 - 18(3) + 6(-1) + 40 = 27 + 3 - 54 - 6 + 40 = 10.$$

Therefore, the minimum value of the expression is 10.

Answer: 10

D.4 Impact of Token-Level Granularity on Entropy Regulation

To validate the necessity of the fine-grained control mechanism proposed in VFPO, we conducted an ablation study by eliminating the token-level variance-aware masking. In this baseline setting (denoted as "Without token level control"), the negative feedback is applied at the sample level without distinguishing between critical and non-critical tokens. Figure 14 illustrates the comparative training dynamics of Group Variance, Group Mean, and Entropy.

As depicted in Figure 14, the absence of token-level granularity leads to suboptimal optimization behaviors:

- **Group Variance and Mean.** The model without token-level control exhibits a rapid decay in Group Variance (teal line), approaching zero early in the training process. Simultaneously, the Group Mean quickly saturates towards 1.0. This indicates that applying constraints solely at the sample level forces the model to converge prematurely to a deterministic policy, reducing the diversity of the generated responses.
- **Entropy Collapse.** A critical observation is the behavior of the policy entropy. The baseline method suffers from severe entropy collapse, where the entropy drops sharply and remains at a negligible level. This suggests that without the selective regulation provided by VFPO, the model becomes overconfident, losing its exploration capability.

In contrast, VFPO (orange/pink line) effectively maintains Group Variance and Entropy at healthy levels throughout the training. By leveraging token-level statistics, VFPO identifies and regulates only the high-variance (critical) tokens while relaxing constraints on low-entropy tokens. This fine-grained mechanism ensures that the negative feedback is targeted precisely, thereby preventing entropy collapse and preserving the model’s generative diversity while optimizing for alignment.

D.5 Efficacy of Difference-Based Clipping Mechanism

In this subsection, we investigate the contribution of our proposed Difference-Based Clipping mechanism through an ablation study. To isolate the effect of this specific component, we compare

the complete VFPO framework against a variant where the Difference-Based Clipping is removed and replaced with the standard PPO-style ratio clipping, while keeping all other VFPO components unchanged.

Standard ratio-based clipping operates on $r_t(\theta) = \pi_{\text{new}}/\pi_{\text{old}}$, which can be disproportionately restrictive for tokens with small absolute probabilities. Even minor absolute changes in low-probability tokens can result in large ratios, triggering the clip and effectively vanishing the gradient. To address this, VFPO introduces a **Difference-Based Clipping** mechanism that constrains the absolute change in probability:

$$\Delta\pi = \text{clip}(\pi_\theta - \pi_{\theta_{\text{old}}}, -\epsilon_{\text{low}}, \epsilon_{\text{high}}). \quad (40)$$

The policy update ratio is subsequently reconstructed as:

$$r_{i,t}(\theta) = \frac{\pi_{\theta_{\text{old}}} + \Delta\pi}{\pi_{\theta_{\text{old}}}}. \quad (41)$$

By imposing constraints on the probability difference $\Delta\pi$ rather than the ratio, this mechanism allows low-probability tokens to update more flexibly, avoiding the "vanishing gradient" trap inherent in ratio-based constraints.

The empirical results of this ablation study are presented in Figure 15. A stark contrast is observed between the two settings: the ablation variant (denoted as "Without New Clip") exhibits a Clip Fraction that collapses to near-zero levels shortly after training begins. This indicates that without our proposed clipping mechanism, the majority of policy updates are aggressively clipped, rendering the optimization inefficient. In contrast, the full VFPO framework maintains the Clip Fraction within a reasonable and active range (approximately 0.10–0.14).

This significant disparity confirms that the sustained Clip Fraction in VFPO is primarily driven by the proposed Difference-Based Clipping mechanism. The results suggest that standard clipping mechanisms are too aggressive for fine-tuning tasks involving long-tail distributions, suppressing updates for tokens that are statistically improbable but reward-critical. Our approach ensures that these tokens retain sufficient gradient magnitude to be effectively "activated" and optimized, thereby facilitating more efficient policy alignment.

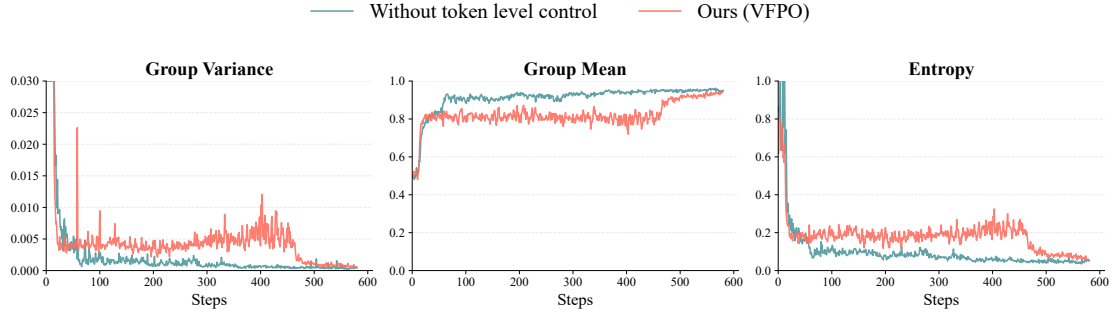


Figure 14: Comparison of optimization dynamics between VFPO and the ablation variant without token-level control. The results demonstrate the evolution of Group Variance, Group Mean, and Entropy during training.

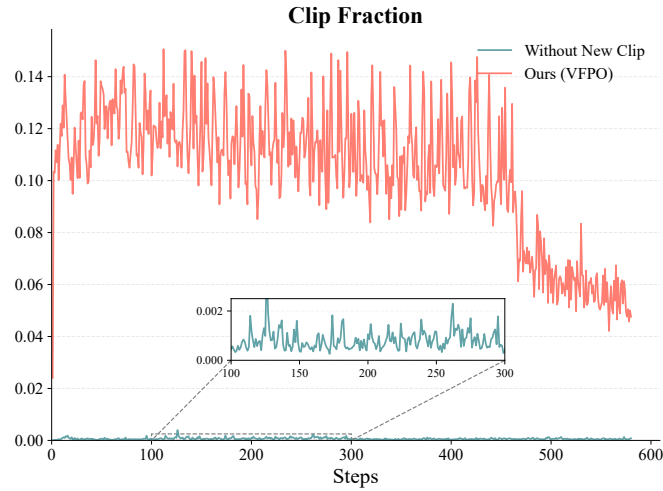


Figure 15: Comparison of the Clip Fraction metric between VFPO (Ours) and the ablation variant without Difference-Based Clipping. The visualization highlights the activation levels of policy updates under different clipping mechanisms.

D.6 Necessity of Sample-Level Reward Modulation

To fully understand the hierarchical contribution of the components within the VFPO framework, we conducted an ablation study specifically targeting the sample-level control mechanism. In this experiment, we eliminated the sample-level reward modulation, effectively forcing the model to rely exclusively on token-level advantage modulation for optimization. We denote this baseline as "Without sample level control".

The objective was to determine whether fine-grained token-level control alone is sufficient to regulate the optimization dynamics, or if a coarse-grained sample-level constraint is requisite for stability.

The empirical results, presented in Figure 16, reveal that token-level control alone is inadequate for maintaining a healthy optimization landscape:

- **Loss of Distributional Stability.** As observed

in the *Group Variance* and *Group Mean* plots, the ablation variant (teal line) exhibits a rapid collapse in variance alongside a premature saturation of the group mean. This indicates that without sample-level modulation to penalize the aggregate behavior of the generated sequences, the model converges too aggressively to a deterministic policy.

- **Entropy Collapse and Reduced Exploration.** The most critical failure mode is observed in the *Entropy* trajectory. The model without sample-level control suffers from severe entropy collapse, dropping significantly below the levels maintained by VFPO (orange/pink line). This sharp decline in entropy signifies that the model has prematurely lost its exploration capability.

Consequently, we conclude that sample-level reward modulation acts as a foundational stabilizer.

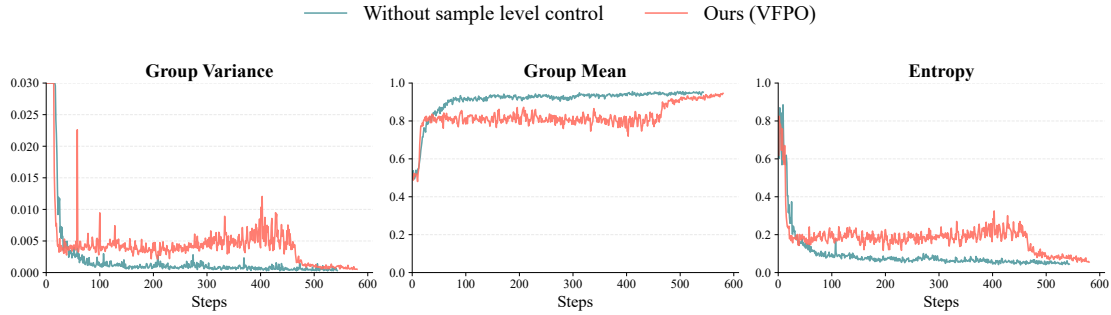


Figure 16: Ablation study on sample-level control. The plots compare the evolution of Group Variance, Group Mean, and Entropy between the full VFPO framework and a variant lacking sample-level reward modulation.

It ensures that the group statistics (mean and variance) are maintained within a reasonable range, thereby allowing the fine-grained token-level mechanism to operate effectively without the system degenerating into a low-entropy, non-exploratory state.

D.7 Necessity of the Annealing Phase for Model Convergence

To validate the design rationale behind the Annealing Phase in the VFPO framework, we conducted an ablation study where this phase was removed. In this experimental setting (denoted as "Without annealing phase"), the variance-aware negative feedback mechanism remains active throughout the entire training process. Consequently, the model is forced to maintain the regulated levels of Group Variance, Group Mean, and Entropy even in the final stages of fine-tuning.

The impact of this ablation on the optimization statistics is illustrated in Figure 17. As expected, the model without the annealing phase maintains a stable Group Mean and keeps the Group Variance and Entropy at elevated levels throughout the training. Unlike the standard VFPO (pink line), which allows these metrics to relax towards the end, the ablated model is prevented from shifting towards a more deterministic distribution.

However, the consequence of this forced stability is evident in the downstream performance. Figure 18 demonstrates that while both methods perform similarly in the early stages, the accuracy of the model without the annealing phase begins to lag behind VFPO in the later stages of training.

This performance degradation can be attributed to the trade-off between exploration and exploitation:

- **Impeded Convergence.** In the late training

phase, the model has typically approached its capability ceiling. At this stage, enforcing high variance (exploration) becomes counter-productive. The model needs to converge towards a confident, low-entropy state to ensure consistent and high-quality outputs.

- **Quality Inconsistency.** By strictly constraining the Group Mean and Variance, the negative feedback mechanism prevents the model from fully exploiting the learned optimal policy. The forced stochasticity introduces unnecessary noise into the generation process, preventing the model from solidifying its knowledge.
- **Optimization Conflict.** High entropy is beneficial for avoiding local optima early on, but it hinders the "polishing" of the policy at the end. The annealing phase in VFPO effectively resolves this by gradually relaxing constraints, allowing the model to transition from diverse exploration to precise exploitation.

In conclusion, the annealing phase is critical to the overall optimization process. It allows the model to leverage the benefits of variance-aware regularization during the exploration phase while ensuring proper convergence and maximizing performance in the final optimization stage.

This is an appendix.

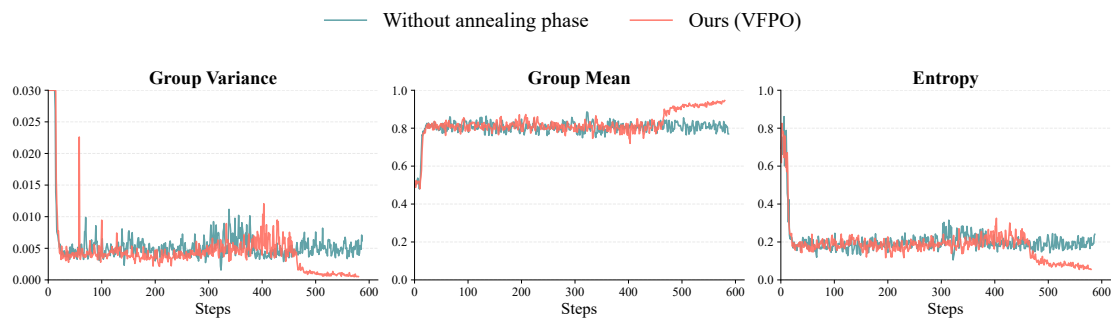


Figure 17: Impact of removing the annealing phase on optimization dynamics. The "Without annealing phase" variant continues to enforce high variance and entropy constraints until the end of training, preventing the natural convergence observed in VFPO.

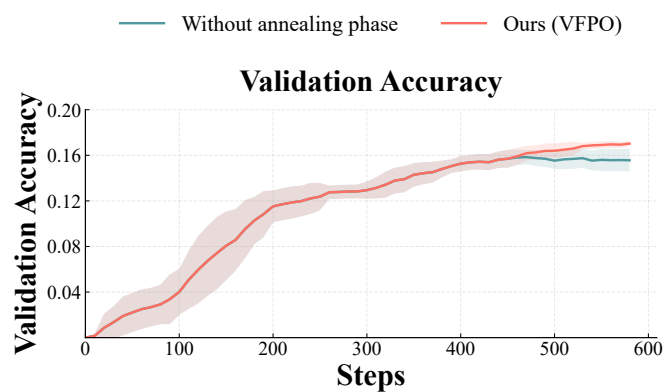


Figure 18: Comparison of Validation Accuracy. The absence of an annealing phase leads to a performance plateau in the late training stages, indicating that forced exploration hinders final convergence.