BEYOND UNIFORM TOKEN-LEVEL TRUST REGIONS IN LLM REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning is widely used to improve reasoning in large language models. A central design choice is how much policy change to allow at each token of a response. Uniform token-level thresholds ignore both token position and preceding policy divergence. Yet early changes can affect longer continuations, and divergence can accumulate along the prefix. To address this, we introduce CPPO (***C**umulative **P**refix Divergence **P**olicy **O**ptimization*), a token-level masking rule that combines position-weighted divergence with a cumulative prefix budget. Divergence at earlier positions receives greater weight, and the prefix budget limits the cumulative weighted divergence. We derive a finite-horizon policy improvement bound under exact token divergence constraints. With a sufficiently small prefix budget, it is tighter than the bound for uniform token-level limits at the same nominal threshold. Experiments on mathematical reasoning show that CPPO outperforms the evaluated baselines across model sizes. Ablation studies support the effectiveness of both components.

$$\mathcal{L}(\theta) = \mathbb{E}_{y\sim\mu}\left[\frac{1}{T}\sum_{t=1}^{|y|} M_t \rho_t \widehat{A}_t\right] \quad M_t = \mathbf{1}\left[\operatorname{sgn}(\widehat{A}_t)(\rho_t - 1) \le 0 \vee I_t\right]$$

PPO $I_t = [\,|\rho_t - 1| \le \epsilon\,]$

DPPO $I_t = [\,D_t \le \delta\,]$

CPPO $I_t = \left[\, w_t D_t \le \delta \wedge \frac{\sum_{j\le t} w_j D_j - \delta}{\sum_{j<t} w_j} \le \delta_b \,\right]$

$$\rho_t = \frac{\pi(y_t|s_t)}{\mu(y_t|s_t)}$$

$$D_t = D(\pi(\cdot|s_t), \mu(\cdot|s_t))$$

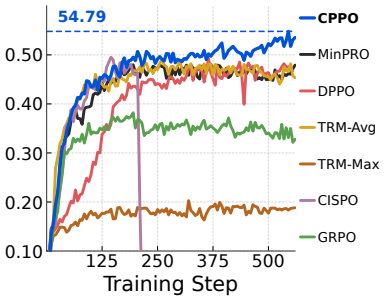

Figure 1: Overview of CPPO. **Left**: token-level masking rules for PPO (Schulman et al., 2017), DPPO (Qi et al., 2026), and CPPO. The prefix fraction applies for $t > 1$; at $t = 1$, the token condition suffices. **Right**: validation AIME24/25/26 Avg@16 on Qwen3-30B-A3B-Base.

1 INTRODUCTION

Reinforcement learning with verifiable rewards (RLVR) has improved mathematical reasoning in large language models (LLMs) by training on responses scored by answer verifiers (Shao et al., 2024; Yu et al., 2025). In a typical training iteration, a rollout policy generates a batch of responses that is reused for several gradient updates. As the current policy changes, those responses can become less representative of what it would generate. This motivates controlling how far the current policy departs from the policy that collected the data, following the trust region approach to policy optimization (Kakade & Langford, 2002; Schulman et al., 2015).

PPO (Schulman et al., 2017) and GRPO (Shao et al., 2024) encourage conservative updates by clipping the sampled-token probability ratio in a surrogate objective. This ratio describes the change in the probability of one observed token; it does not by itself specify the divergence between the two policies' distributions over the vocabulary. Distribution-based approaches instead compare the rollout and current policies' next-token distributions at the same sampled prefix. This token-level

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

divergence can guide token masking (Qi et al., 2026) or the projection of distributions used in the objective (Becker et al., 2026). Reduced representations make these comparisons practical for large vocabularies.

The choice of divergence statistic, however, still leaves open how much change should be allowed at different points in a response. Under a uniform token-level threshold, the divergence test is identical near the beginning and near the end. These positions have different remaining horizons in autoregressive generation. A change in an early conditional distribution can affect a long continuation, whereas a change at the final position affects only the last token. This motivates assigning greater weight to earlier divergences when controlling policy change.

The current divergence score also leaves out changes at preceding positions. Consider two prefixes at the same position with identical current divergence scores. One may follow little change in earlier conditional distributions, while the other follows sustained deviations over many tokens. A test based only on the current score gives the same answer in both cases. Yet preceding conditional changes can alter the probability of reaching a prefix, even when the current next-token distributions remain close. A cumulative budget can distinguish these histories and reduce the remaining allowance after sustained deviations.

We address this allocation problem with CPPO (***C**umulative **P**refix Divergence **P**olicy **O**ptimization*). Figure 1 summarizes the rule and representative validation curves. Our main contributions are:

- We combine position-weighted divergence and a cumulative prefix budget in a token-level mask. The mask gives greater weight to divergence at earlier positions and uses preceding weighted divergences to determine the current allowance (Section 3).
- We derive a finite-horizon policy improvement bound under exact token divergence constraints. At the same nominal threshold, prefix control gives a tighter bound than uniform token-level limits when the prefix budget is sufficiently small (Section 4).
- We evaluate CPPO against six baselines on AIME24/25/26 across four Qwen3 configurations, from 1.7B to 30B-A3B. Within the same number of training steps, CPPO improves best validation mean $\mathrm{Avg@16}$ over the strongest baseline in every setting. Gains range from 0.91 to 5.56 percentage points, with the largest gain on Qwen3-30B-A3B-Base. Ablation studies support the effectiveness of our design choices (Section 5).

2 PRELIMINARIES

Given a prompt $x \sim \mathcal{D}$, a language model policy $\pi = \pi_\theta$ generates a response $y = (y_1, \ldots, y_N)$ autoregressively, with state $s_t = (x, y_{<t})$ and length $N \leq T$, where $T \geq 2$ is the maximum response horizon. A verifier assigns a reward $R(x, y)$, and training seeks to maximize $J(\pi) = \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi(\cdot|x)}[R(x, y)]$.

Policy gradient updates. Training replays responses from a rollout policy μ while updating π_θ. With advantage estimates $\hat{A}_t$, we maximize the surrogate $\mathcal{L}_\mu(\pi_\theta) = \mathbb{E}_\mu[T^{-1} \sum_{t=1}^{N} \rho_t \hat{A}_t]$. Its gradient and sampled-token ratio are (Schulman et al., 2017)

$$\nabla_\theta \mathcal{L}_\mu(\pi_\theta) = \mathbb{E}_\mu \left[\frac{1}{T} \sum_{t=1}^{N} \rho_t \hat{A}_t \nabla_\theta \log \pi_\theta(y_t \mid s_t) \right], \qquad \rho_t = \frac{\pi_\theta(y_t \mid s_t)}{\mu(y_t \mid s_t)}. \tag{1}$$

Here $\mathbb{E}_\mu$ averages over prompts and rollout responses, with μ and $\hat{A}_t$ held fixed. Each token's contribution favors increasing its probability for $\hat{A}_t > 0$ and decreasing it for $\hat{A}_t < 0$. We use GRPO advantages (Shao et al., 2024): $\hat{A}_{i,t} = (R_i - \bar{R})/\sigma_R$, shared by all tokens in response i, where $\bar{R}$ and σ_R are the mean and standard deviation of rewards for responses to the same prompt. Zero-variance groups are skipped.

Token-level masks. To control policy change, PPO tests the sampled-token ratio, whereas DPPO tests the divergence between next-token distributions (Qi et al., 2026). We use total variation (TV), $D_t = \frac{1}{2} \sum_{a\in\mathcal{V}} |\pi(a \mid s_t) - \mu(a \mid s_t)|$, where $\mathcal{V}$ is the vocabulary, and its Top-K reduced

approximation $\widehat{D}_t \leq D_t$ (Appendix C.2). For thresholds $\epsilon, \delta > 0$, the masks are

$$M_t^{\mathrm{PPO}} = \mathbf{1}\left[\hat{A}_t(\rho_t - 1) \leq 0 \ \vee \ |\rho_t - 1| \leq \epsilon\right], \tag{2}$$

$$M_t^{\mathrm{uniform}} = \mathbf{1}\left[\hat{A}_t(\rho_t - 1) \leq 0 \ \vee \ \widehat{D}_t \leq \delta\right]. \tag{3}$$

The mask multiplies each token's gradient contribution: $M_t = 1$ retains it and $M_t = 0$ removes it. The condition $\hat{A}_t(\rho_t - 1) > 0$ identifies *outward* contributions that favor moving the sampled-token ratio farther from one; only these are subject to the threshold test. The uniform rule uses the same δ at every position, independently of preceding divergence.

3 CPPO

CPPO allocates divergence allowance using token position and preceding divergence. Given a rollout response, its inputs are the token ratios ρ_t, advantages $\hat{A}_t$, and divergence scores $\widehat{D}_t$ between the current policy π and rollout policy μ. We first weight each score by position, then reduce its allowance according to the divergence accumulated along the sampled prefix. The resulting mask selects contributions to the policy gradient.

3.1 Position-weighted divergence

An early change in a conditional token distribution can affect a longer continuation than a change near the end of a response. We assign larger weights to earlier divergences in the masking rule. For the fixed response horizon T, we use a decreasing linear weight with floor $0 < w_{\min} \leq 1$:

$$w_t = 1 - \frac{1 - w_{\min}}{T - 1}(t - 1), \qquad w_t \widehat{D}_t \leq \delta. \tag{4}$$

The equivalent threshold on $\widehat{D}_t$, δ / w_t, ranges from δ at the first token to $\delta / w_{\min}$ at the fixed horizon T. Thus, the same divergence score can fail the test near the start of a response and pass it later. Responses that end before T follow the same schedule. The weights act on divergence scores, leaving the advantage estimates unchanged. The positive floor keeps the threshold finite at T.

3.2 Cumulative prefix budget

Position weighting alone sets a threshold independently at each token. Changes at preceding positions can alter the probability of reaching the sampled prefix even when the current next-token distributions remain close. To account for these changes, we track the cumulative weighted divergence and cumulative weight:

$$S_t = \sum_{j=1}^{t} w_j \widehat{D}_j, \qquad W_t = \sum_{j=1}^{t} w_j, \qquad S_0 = W_0 = 0. \tag{5}$$

Every valid token enters these sums, including masked tokens. Masking removes a token's gradient contribution; it does not change the divergence measured at that position.

For a budget rate $\delta_b > 0$, the prefix budget condition is

$$S_t \leq \delta + \delta_b W_{t-1}. \tag{6}$$

The budget starts at δ, so the first token has its full allowance. It then grows with the cumulative weight of the preceding positions at a rate set by δ_b.

Subtracting S_{t-1} from the prefix budget gives the remaining prefix allowance. We combine this with the token threshold δ and denote the resulting allowance for $w_t \widehat{D}_t$ by b_t:

$$b_t = \min\{\delta, \, \delta + \delta_b W_{t-1} - S_{t-1}\}, \qquad w_t \widehat{D}_t \leq b_t. \tag{7}$$

The rate δ_b controls budget growth, while b_t is computed at each position from the preceding prefix. The allowance b_t stays at δ as long as $S_{t-1} \leq \delta_b W_{t-1}$. Above this level, the excess weighted divergence is subtracted from the current token's allowance. Dividing by w_t gives the effective

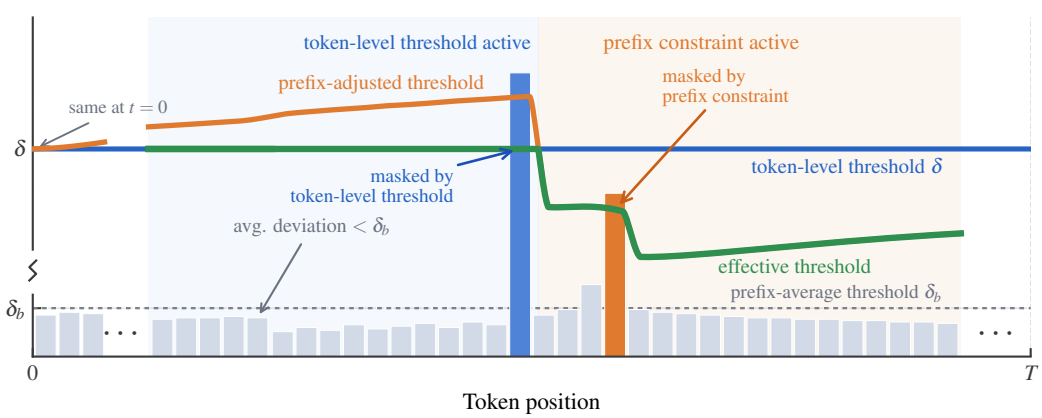

Figure 2: Illustration of the CPPO divergence conditions. Bars show schematic weighted scores. The blue and orange curves give the token-level limit δ and remaining prefix allowance; their minimum (green) is the current allowance b_t. Highlighted bars exceed this allowance and illustrate tokens whose outward updates are masked.

threshold b_t/w_t on $\widehat{D}_t$. If $b_t < 0$, no nonnegative score passes the divergence test. When subsequent scores fall below δ_b, the remaining allowance can recover as the budget grows.

For a schematic example, take $\delta = 0.15$, $\delta_b W_{t-1} = 0.10$, and a current weighted score $w_t \widehat{D}_t = 0.09$. A prefix sum of $S_{t-1} = 0.08$ gives $b_t = 0.15$, so the score passes; $S_{t-1} = 0.18$ gives $b_t = 0.07$, so the same score fails. Figure 2 illustrates this dependence on preceding divergence: the tall blue bar reduces the subsequent allowance, causing the later orange bar to fail the prefix test despite lying below the token-level limit δ.

3.3 Token Mask and Surrogate Objective

CPPO retains the same directional condition as uniform masking: only outward contributions must pass the divergence test. Combining this condition with the allowance above gives

$$M_t^{\mathrm{CPPO}} = \mathbf{1}\left[\hat{A}_t(\rho_t - 1) \le 0 \ \vee \ w_t \widehat{D}_t \le b_t\right]. \tag{8}$$

We maximize the masked surrogate

$$\mathcal{L}_\mu^{\mathrm{CPPO}}(\pi) = \mathbb{E}_{x\sim\mathcal{D},\, y\sim\mu(\cdot|x)}\left[\frac{1}{T}\sum_{t=1}^{N} \mathrm{sg}\big(M_t^{\mathrm{CPPO}}\big)\, \rho_t \hat{A}_t\right], \tag{9}$$

where sg stops gradients through the mask, and the advantage estimates are held fixed during optimization. Its gradient is Equation (1) with each token contribution multiplied by M_t^{CPPO}. The normalization remains $1/T$ regardless of how many tokens are masked. Shared parameters allow retained gradients to change probabilities at masked positions, so masking does not freeze those probabilities or enforce a hard constraint on the updated policy.

Relative to a uniform threshold δ, CPPO can both retain additional tokens and mask tokens that the uniform rule would retain. With an inactive prefix budget, $b_t/w_t = \delta/w_t \ge \delta$; accumulated divergence can instead reduce b_t/w_t below δ.

At each loss evaluation, we recompute divergence scores against the fixed rollout policy and form the cumulative sums within each response. These sums are not carried across optimizer steps. The budget rate δ_b can be prescribed in advance or calibrated separately for each response. For the latter, we take the 90th percentile of divergence scores over valid tokens in the response and clamp it to $[\delta_b^{\min}, 2\delta_b^{\min}]$, where $\delta_b^{\min} > 0$ is a chosen minimum. In both cases, the rate is held fixed across positions during a mask evaluation. Calibration uses the completed response; the sum S_{t-1} uses only preceding positions. Appendix C.3 gives pseudocode, the batched implementation, and a soft mask variant.

4 THEORETICAL ANALYSIS

We bound the error of a token-level policy surrogate using exact TV. The bound gives larger coefficients to earlier divergences. Cumulative prefix constraints tighten the bound under the budget conditions below. Proofs are in Appendix B.

4.1 FINITE-HORIZON PERFORMANCE DIFFERENCE IDENTITY

Consider a horizon $T \geq 2$ and policies μ, π with common support. Assume that $|R(x, y)| \leq \xi$ for some $\xi > 0$. Terminated responses are extended to the fixed horizon with an absorbing state where both policies agree. The suffix likelihood ratio is

$$\rho_{t+1:T} = \prod_{j=t+1}^{T} \rho_j, \qquad \rho_{T+1:T} = 1,$$

using the sampled-token ratios from Section 2. Autoregressive factorization gives the following decomposition (Qi et al., 2026):

$$\begin{aligned} J(\pi) - J(\mu) &= L'_\mu(\pi) - \Delta(\mu, \pi), \\ L'_\mu(\pi) &:= \mathbb{E}_\mu\left[R(x, y) \sum_{t=1}^{T} (\rho_t - 1)\right], \\ \Delta(\mu, \pi) &:= \mathbb{E}_\mu\left[R(x, y) \sum_{t=1}^{T} (\rho_t - 1)(1 - \rho_{t+1:T})\right]. \end{aligned} \tag{10}$$

Here $\mathbb{E}_\mu$ includes the prompt distribution. The surrogate L'_μ sums reward-weighted changes in individual token ratios, while Δ captures their interactions with changes in the suffix distribution (Lemma 2). We control this error by bounding the conditional divergences.

4.2 POSITION-DEPENDENT SURROGATE ERROR BOUND

Let d_t^μ be the rollout prefix distribution. Define the expected divergence and deterministic token-level limits by

$$u_t = \mathbb{E}_{s_t \sim d_t^\mu}[D_t], \qquad D_t \leq \ell_t, \qquad \bar{\ell} = \max_t \ell_t.$$

Assume these limits hold μ-almost surely. Bounding suffix divergence by the sum of the remaining token-level limits gives (Proposition 4)

$$\begin{aligned} |\Delta(\mu, \pi)| &\leq 4\xi \sum_{t=1}^{T-1} u_t \sum_{j=t+1}^{T} \ell_j \leq \sum_{t=1}^{T-1} \lambda_t u_t, \\ \lambda_t &:= 4\xi \bar{\ell}(T - t). \end{aligned} \tag{11}$$

Each expected divergence u_t is weighted by the limits on the suffix that follows it. Earlier divergences therefore enter the bound with larger coefficients, motivating the decreasing weights in Equation (4). The bound retains the expected divergences at individual positions rather than replacing each with a common token limit. This lets us use the prefix constraints to control their accumulated contribution to the surrogate error.

4.3 POLICY IMPROVEMENT BOUND UNDER PREFIX CONSTRAINTS

To examine prefix control, fix positive weights w_t, weighted token limits $c_t > 0$, and a budget rate $\delta_b > 0$. Define

$$P_m = \sum_{j=1}^{m} w_j D_j, \qquad \ell_t = \frac{c_t}{w_t}.$$

Here P_m is the exact TV counterpart of S_m, and ℓ_t is the unweighted token limit. With the cumulative weights from Equation (5), impose

$$\begin{aligned} w_t D_t &\leq c_t, \qquad t = 1, \ldots, T, \\ P_m &\leq \delta_b W_m, \quad m = 1, \ldots, T - 1. \end{aligned} \tag{12}$$

Requiring the budget at every prefix limits early accumulation even when later divergences are small. A bound on the full response average alone can hide such an early excess.

Theorem 1 (Policy improvement bound under prefix constraints)**.** *Suppose Equation* (12) *holds μ-almost surely and the coefficient ratios λ_t/w_t are non-increasing for $t = 1, \ldots, T-1$, with λ_t defined in Equation* (11)*. Then*

$$J(\pi) - J(\mu) \geq L'_\mu(\pi) - 2\xi T(T-1)\bar{\ell}\delta_b. \tag{13}$$

For a constant weighted token limit $c_t \equiv \delta$ and weights in $[w_{\min}, 1]$, this gives

$$J(\pi) - J(\mu) \geq L'_\mu(\pi) - 2\xi T(T-1)\delta\delta_b/w_{\min}. \tag{14}$$

The proof in Appendix B.4 uses Abel summation to convert the prefix constraints into a bound on the weighted divergence sum in Equation (11). The condition that λ_t/w_t is non-increasing makes the coefficients of the prefix slacks nonnegative in this summation. The linear schedule satisfies this condition (Proposition 5).

For comparison, a uniform exact TV limit $0 < \delta \leq 1$ at every token, without prefix constraints, gives

$$D_t \leq \delta \text{ for all } t \quad \Longrightarrow \quad |\Delta(\mu, \pi)| \leq 2\xi T(T-1)\delta^2.$$

The residual term in Equation (14) is smaller when $\delta_b < \delta w_{\min}$. This compares the bounds under a shared nominal threshold; position weighting changes the effective token limits. Corollaries 6 and 7 give both this comparison and the case with matched unweighted limits. For long horizons, a bound linear in T can be smaller than the quadratic bound above. Appendix B.9 derives this bound for both constraint sets. When each uses the smaller of its two bounds, prefix control still improves the comparison under the same condition $\delta_b < \delta w_{\min}$.

4.4 Initial allowance and the CPPO update

CPPO includes an initial allowance δ in its prefix budget (Equation (6)). If the corresponding token and prefix conditions hold μ-almost surely with exact TV, then for the linear schedule, Proposition 8 gives

$$|\Delta(\mu, \pi)| \leq 2\xi T(T-1)\bar{\ell}\delta_b + 4\xi\bar{\ell}(T-1)\delta. \tag{15}$$

The second term accounts for the initial allowance; its ratio to the leading term is $2\delta/(T\delta_b)$. In the linear bound, the added term is at most $4\xi\delta/w_{\min}$, independently of T. Appendix B.9 compares both corrected bounds with uniform token limits.

For efficient computation, CPPO uses reduced TV to approximate full-vocabulary TV, following Qi et al. (2026). The bounds describe surrogate error for policies satisfying the stated divergence constraints.

5 Experiments

We evaluate CPPO on mathematical reasoning across four Qwen3 configurations, then examine its components and design choices.

5.1 Experimental setup

Models and training. We train on roughly 17k verifiable mathematical reasoning prompts from DAPO-Math-17k (Yu et al., 2025). We use the `verl` (Sheng et al., 2025) GRPO/DAPO training stack. The four Qwen3 (Yang et al., 2025) configurations are Qwen3-1.7B (post-trained), Qwen3-1.7B-Base, Qwen3-8B-Base, and Qwen3-30B-A3B-Base. The dense models use an 8k rollout horizon and eight responses per prompt; the 30B-A3B model uses 16k and sixteen responses. Training data and rollout configurations are matched across methods within each model setting.

Evaluation protocol. We measure $\mathrm{Avg@16}$, the average correctness over sixteen sampled responses per problem, on AIME24, AIME25, and AIME26. The reported score is the unweighted mean across these benchmarks. For each method, we select the checkpoint with the highest mean validation score within the same training-step window $[0, T^{\mathrm{stop}}]$ for that model. These are best validation results;

Table 1: Best validation AIME24/25/26 Avg@16 across four Qwen3 configurations (%, higher is better), selected within a matched training-step window for each model. **Bold** and underline mark the best and second-best reported scores. *Collapse* marks the CISPO run that diverged at step 215. Table 4 gives the window endpoints and per-benchmark scores at the selected checkpoints.

Method	1.7B	1.7B-Base	8B-Base	30B-A3B-Base
GRPO	27.91	8.89	23.96	38.19
MinPRO	27.71	11.04	<u>29.72</u>	48.12
CISPO	<u>28.82</u>	<u>11.87</u>	29.58	*collapse*
DPPO	28.19	10.90	28.89	<u>49.23</u>
TRM-Max	25.21	9.72	26.73	20.27
TRM-Avg	26.87	11.70	27.98	48.96
CPPO (ours)	**31.88**	**12.78**	**31.11**	**54.79**

the three benchmark scores come from the same selected checkpoint. Appendix C.1 provides the decoding and training settings; Table 4 gives the per-benchmark results.

Baselines. GRPO (Shao et al., 2024) and CISPO (Chen et al., 2025) operate on sampled-token ratios, with CISPO using asymmetric clip thresholds. MinPRO (Lei et al., 2026) uses a prefix ratio surrogate without a cumulative distributional budget. TRM-Max and TRM-Avg (Li et al., 2025) mask entire responses using maximum and mean token-level KL, respectively. The GRPO row uses the Clip-Higher setting $(\epsilon_{\text{low}}, \epsilon_{\text{high}}) = (0.2, 0.28)$. The other baseline hyperparameters follow their original papers. DPPO (Qi et al., 2026) uses a uniform token-level divergence threshold. It shares the Top-K (K=20) reduced TV score and nominal threshold scale δ with CPPO.

CPPO configuration. We set $\delta = 0.15$ for the dense models, $\delta = 0.2$ for 30B-A3B-Base, and $w_{\min} = 0.8$ throughout the main experiments. Qwen3-1.7B uses a fixed budget rate $\delta_b = 0.015$. For Base models, we adapt the rate to each response by taking the 90th percentile of divergence scores over its valid tokens and clamping it to $[\delta_b^{\min}, 2\delta_b^{\min}]$, with $\delta_b^{\min} = 0.02$. Appendix Figures 6 and 7 provide position and budget diagnostics.

5.2 Main results

CPPO achieves the highest mean validation score in all four configurations (Table 1), with gains over the strongest baseline ranging from 0.91 to 5.56 percentage points. The largest gain is on Qwen3-30B-A3B-Base.

The gains over DPPO hold with the divergence score and nominal threshold matched. This comparison changes both budget allocation and effective thresholds; Section 5.3 examines the components separately.

Figure 3 shows that CPPO's advantage on the Base models is clearest in later training; the methods often overlap earlier. On 30B-A3B-Base, CPPO continues to improve after several baselines plateau, while CISPO collapses. Appendix C.5 gives the corresponding training reward, response length, and log probability diagnostics.

5.3 Ablation studies

We examine position weighting, the prefix budget, the ordering of position weights, and the masking rule on Qwen3-1.7B (Figure 4).

Position weighting and prefix budget. Figure 4 (left) removes one component at a time. CPPO w/o Position Weight retains the prefix budget with $w_t \equiv 1$; CPPO w/o Prefix Budget retains only the position-weighted divergence threshold. Removing either component lowers the best validation score, although both variants still outperform the uniform token-level baseline. Appendix C.6 presents three response excerpts illustrating the mask decisions.

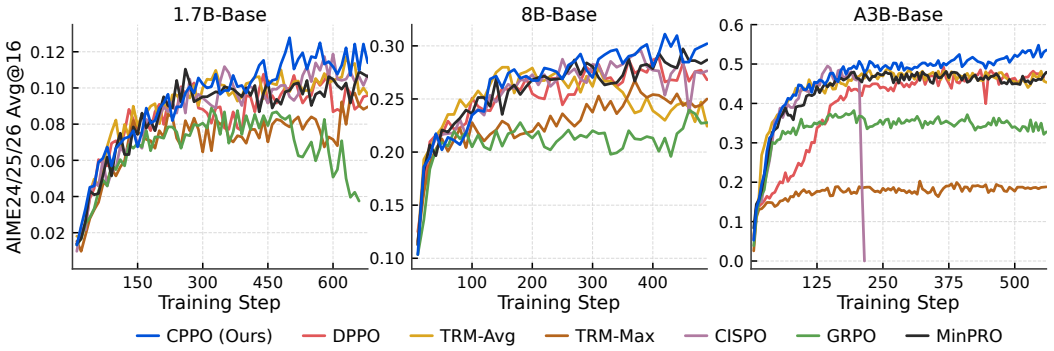

Figure 3: Validation performance during training on Qwen3-1.7B-Base, Qwen3-8B-Base, and Qwen3-30B-A3B-Base (left to right). Each curve shows the mean AIME24/25/26 Avg@16 against training steps within the model's comparison window. The CISPO run in the right panel collapses at step 215.

Ordering of position weights. To test the benefit of assigning larger weights to earlier divergences, we compare the ordered schedule with a shuffled version (Figure 4, middle). Shuffling retains every weight value but randomly reassigns the values to positions. Both variants omit the prefix budget, so the comparison isolates the ordering of divergence weights. The ordered schedule achieves higher scores late in training, whereas the shuffled variant remains close to uniform masking.

Hard versus soft masking. Figure 4 (right) replaces the hard mask with a non-increasing soft gate that attenuates token gradients beyond the budget boundary. Both variants yield similar validation trajectories; we retain the hard mask as the default. Appendix C.3 specifies the soft gate.

5.4 Hyperparameter sensitivity and divergence variants

Figure 5 evaluates the budget settings and divergence score on Qwen3-1.7B-Base.

Budget minimum and weight floor. Figure 5 (left) varies the budget parameter and weight floor under the same calibration rule. Relative to the default setting, we increase $\delta_b^{\min}$ from 0.02 to 0.03 or reduce $w_{\min}$ from 0.8 to 0.6. The setting with the larger budget minimum has lower scores early in training, but the three configurations reach similar scores later.

Divergence measure and masking granularity. Figure 5 (middle) compares the default TV score with KL. The KL variant uses TRM's threshold values, $\delta = 0.1$ and $\delta_b = 0.002$, within the CPPO rule. It reaches scores comparable to the TV variant and above the TRM Max&Avg control. The control uses the same KL thresholds to mask whole responses by their maximum and mean divergence.

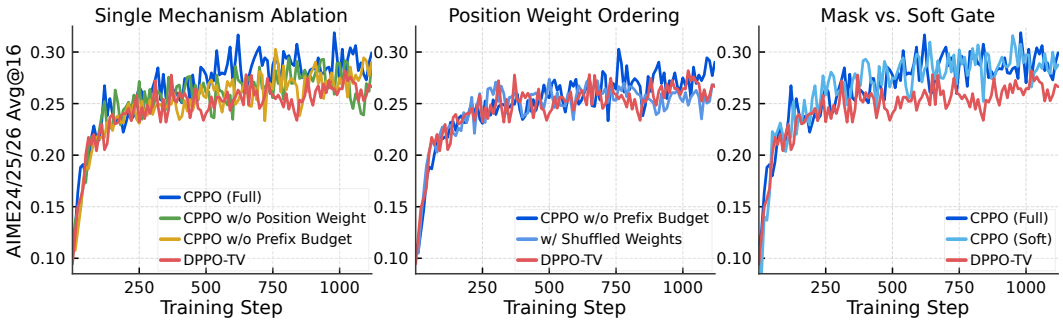

Figure 4: Ablations on Qwen3-1.7B, reporting mean validation AIME24/25/26 Avg@16 over training steps. **Left**: remove position weighting or the prefix budget. **Middle**: compare ordered and shuffled weights with the prefix budget removed in both variants. **Right**: compare hard and soft masking with both components. Uniform token-level masking (DPPO-TV) is included in each panel.

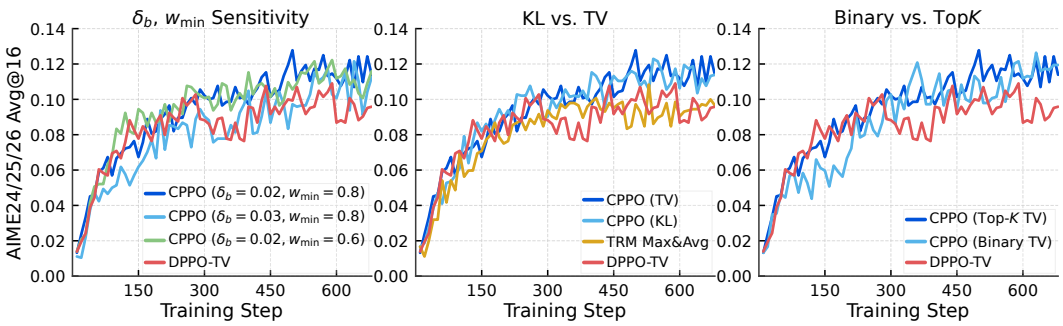

Figure 5: Sensitivity analyses on Qwen3-1.7B-Base, using mean validation AIME24/25/26 Avg@16. **Left**: sensitivity to the budget parameter and weight floor. **Middle**: TV and KL within CPPO, with a sequence-level KL masking control (TRM Max&Avg). **Right**: Top-K and Binary TV approximations.

Under these thresholds, cumulative prefix aggregation with token-level masking performs better than the response-level control.

TV approximation. Figure 5 (right) compares Binary and Top-K TV (Qi et al., 2026) with the CPPO rule fixed. Binary TV has lower validation scores early in training but reaches similar scores later. CPPO improves over uniform masking with both approximations.

6 Related Work

Likelihood ratio policy optimization. PPO (Schulman et al., 2017) and GRPO (Shao et al., 2024) use a clipped token ratio surrogate; DAPO (Yu et al., 2025) raises the upper clipping threshold, and CISPO (Chen et al., 2025) clips the importance weights used in the gradient. GSPO uses a length-normalized ratio for the full response (Zheng et al., 2025), while MinPRO weights the current ratio by the minimum ratio in the preceding prefix (Lei et al., 2026). MinPRO uses prefix information to reweight the surrogate; CPPO uses accumulated divergence to set the allowance for each token update.

Distributional trust regions. Classical trust region methods constrain policy divergence to bound surrogate error (Kakade & Langford, 2002; Schulman et al., 2015; Peters et al., 2010). DPPO (Qi et al., 2026) replaces the ratio test for the sampled token with a uniform token-level divergence threshold; TROLL (Becker et al., 2026) projects sparse next-token distributions into a KL trust region; and TRM (Li et al., 2025) uses response-level divergence statistics to mask whole responses. CPPO retains token-level masking and allocates divergence allowance using both position and preceding divergence. Its prefix budget can mask a token with small local divergence when earlier deviations have used up the available allowance. Appendix A compares the statistics and update rules in more detail and discusses related work on token weighting and learning dynamics.

7 Conclusion

We introduced CPPO to allocate divergence allowance across a response using position-weighted divergence and a cumulative prefix budget. Together, these components make token updates depend on both where divergence occurs and how much has already accumulated. Our analysis relates surrogate error to the remaining horizon. Under the stated exact TV constraints and budget conditions, prefix control yields a tighter bound than uniform token limits. Across four Qwen3 configurations, CPPO outperforms all evaluated baselines in best validation scores on mathematical reasoning benchmarks. Ablations support both components and the use of ordered position weights.

AI USE STATEMENT

We used generative AI tools to aid and polish the writing of this paper, specifically to edit wording and paper organization for readability and to help with LaTeX formatting. We also used generative AI tools for retrieval and discovery, to search for and identify related work. We also used these tools to help revise and check mathematical derivations and develop scripts for diagnostic analysis and plotting. We reviewed and revised all AI-assisted text and verified each AI-retrieved reference against the original publication before citing it. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Abbas Abdolmaleki, Jost Tobias Springenberg, Yuval Tassa, Rémi Munos, Nicolas Heess, and Martin A. Riedmiller. Maximum a posteriori policy optimisation. *CoRR*, abs/1806.06920, 2018. URL http://arxiv.org/abs/1806.06920.

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International conference on machine learning*, pp. 22–31. Pmlr, 2017.

Marcin Andrychowicz, Anton Raichuk, Piotr Stańczyk, Manu Orsini, Sertan Girgin, Raphaël Marinier, Leonard Hussenot, Matthieu Geist, Olivier Pietquin, Marcin Michalski, Sylvain Gelly, and Olivier Bachem. What matters for on-policy deep actor-critic methods? a large-scale study. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=nIAxjsniDzg.

Philipp Becker, Niklas Freymuth, Serge Thilges, Fabian Otto, and Gerhard Neumann. TROLL: Trust regions improve reinforcement learning for large language models. In *International Conference on Learning Representations*, pp. 53096–53135, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/56fbf5a2109a6c17372209c9fa698857-Paper-Conference.pdf.

Aili Chen, Aonian Li, Bangwei Gong, Binyang Jiang, Bo Fei, Bo Yang, Boji Shan, Changqing Yu, Chao Wang, Cheng Zhu, Chengjun Xiao, Chengyu Du, Chi Zhang, Chu Qiao, Chunhao Zhang, Chunhui Du, Congchao Guo, Da Chen, Deming Ding, Dianjun Sun, Dong Li, Enwei Jiao, Haigang Zhou, Haimo Zhang, Han Ding, Haohai Sun, Haoyu Feng, Huaiguang Cai, Haichao Zhu, Jian Sun, Jiaqi Zhuang, Jiaren Cai, Jiayuan Song, Jin Zhu, Jingyang Li, Jinhao Tian, Jinli Liu, Junhao Xu, Junjie Yan, Junteng Liu, Junxian He, Kaiyi Feng, Ke Yang, Kecheng Xiao, Le Han, Leyang Wang, Lianfei Yu, Liheng Feng, Lin Li, Lin Zheng, Linge Du, Lingyu Yang, Lunbin Zeng, Minghui Yu, Mingliang Tao, Mingyuan Chi, Mozhi Zhang, Mujie Lin, Nan Hu, Nongyu Di, Peng Gao, Pengfei Li, Pengyu Zhao, Qibing Ren, Qidi Xu, Qile Li, Qin Wang, Rong Tian, Ruitao Leng, Shaoxiang Chen, Shaoyu Chen, Shengmin Shi, Shitong Weng, Shuchang Guan, Shuqi Yu, Sichen Li, Songquan Zhu, Tengfei Li, Tianchi Cai, Tianrun Liang, Weiyu Cheng, Weize Kong, Wenkai Li, Xiancai Chen, Xiangjun Song, Xiao Luo, Xiao Su, Xiaobo Li, Xiaodong Han, Xinzhu Hou, Xuan Lu, Xun Zou, Xuyang Shen, Yan Gong, Yan Ma, Yang Wang, Yiqi Shi, Yiran Zhong, and Yonghong Duan. MiniMax-M1: Scaling test-time compute efficiently with lightning attention. *CoRR*, abs/2506.13585, June 2025. URL https://doi.org/10.48550/arXiv.2506.13585.

Xiangxiang Chu, Hailang Huang, Xiao Zhang, Fei Wei, and Yong Wang. GPG: A simple and strong reinforcement learning baseline for model reasoning. *CoRR*, abs/2504.02546, April 2025. URL https://doi.org/10.48550/arXiv.2504.02546.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, Zhiyuan Liu, Hao Peng, LEI BAI, Wanli Ouyang, Yu Cheng, Bowen Zhou, and Ning Ding. The entropy mechanism of reinforcement learning for reasoning language models, 2025. URL https://openreview.net/forum?id=vXoksdcfqC.

Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Firdaus Janoos, Larry Rudolph, and Aleksander Madry. Implementation matters in deep RL: A case study on PPO and TRPO. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=r1etN1rtPB.

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

Jian Hu, Jason Klein Liu, Haotian Xu, and Wei Shen. REINFORCE++: Stabilizing critic-free policy optimization with global advantage normalization. *arXiv preprint arXiv:2501.03262*, 2025.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the nineteenth international conference on machine learning*, pp. 267–274, 2002.

Shiye Lei, Zhihao Cheng, and Dacheng Tao. A step back: Prefix importance ratio stabilizes policy optimization. *arXiv preprint arXiv:2601.22718*, 2026.

Yingru Li, Jiacai Liu, Jiawei Xu, Yuxuan Tong, Ziniu Li, Qian Liu, and Baoxiang Wang. Trust region masking for long-horizon LLM reinforcement learning. *arXiv preprint arXiv:2512.23075*, 2025.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding R1-Zero-like training: A critical perspective. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=5PAF7PAY2Y.

Chiyu Ma, Shuo Yang, Kexin Huang, Jinda Lu, Haoming Meng, Shangshang Wang, Bolin Ding, Soroush Vosoughi, Guoyin Wang, and Jingren Zhou. FIPO: Eliciting deep reasoning with Future-KL influenced policy optimization. *arXiv preprint arXiv:2603.19835*, 2026.

Jan Peters, Katharina Mulling, and Yasemin Altun. Relative entropy policy search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 24, pp. 1607–1612, 2010.

Penghui Qi, Xiangxin Zhou, Zichen Liu, Tianyu Pang, Chao Du, Min Lin, and Wee Sun Lee. Rethinking the trust region in LLM reinforcement learning. *arXiv preprint arXiv:2602.04879*, 2026.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897. PMLR, 2015.

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, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300, 2024. URL https://doi.org/10.48550/arXiv.2402.03300.

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*, EuroSys '25, pp. 1279–1297, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400711961. doi: 10.1145/3689031.3696075. URL https://doi.org/10.1145/3689031.3696075.

H. Francis Song, Abbas Abdolmaleki, Jost Tobias Springenberg, Aidan Clark, Hubert Soyer, Jack W. Rae, Seb Noury, Arun Ahuja, Siqi Liu, Dhruva Tirumala, Nicolas Heess, Dan Belov, Martin A. Riedmiller, and Matthew M. Botvinick. V-MPO: On-policy maximum a posteriori policy optimization for discrete and continuous control. *CoRR*, abs/1909.12238, 2019. URL http://arxiv.org/abs/1909.12238.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for LLM reasoning. *CoRR*, abs/2506.01939, June 2025. URL https://doi.org/10.48550/arXiv.2506.01939.

Yuhui Wang, Hao He, and Xiaoyang Tan. Truly proximal policy optimization. In *UAI*, pp. 113–122, 2019a. URL http://proceedings.mlr.press/v115/wang20b.html.

Yuhui Wang, Hao He, Xiaoyang Tan, and Yaozhong Gan. Trust region-guided proximal policy optimization. *CoRR*, abs/1901.10314, 2019b. URL http://arxiv.org/abs/1901.10314.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jian Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *CoRR*, abs/2505.09388, May 2025. URL https://doi.org/10.48550/arXiv.2505.09388.

Deokgyu Yoon, Hyungkyu Kang, Joongkyu Lee, Byeongchan Kim, Gyungin Shin, Sungrae Park, and Min-hwan Oh. Multi-step likelihood-ratio correction for reinforcement learning with verifiable rewards. *arXiv preprint arXiv:2605.20865*, 2026.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, YuYue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, LingJun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Yonghui Wu, and Mingxuan Wang. DAPO: An open-source LLM reinforcement learning system at scale. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=2a36EMSSTp.

Yuzhong Zhao, Yue Liu, Junpeng Liu, Jingye Chen, Xun Wu, Yaru Hao, Tengchao Lv, Shaohan Huang, Lei Cui, Qixiang Ye, Fang Wan, and Furu Wei. Geometric-mean policy optimization. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=nCEs0tSwc2.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group sequence policy optimization. *CoRR*, abs/2507.18071, July 2025. URL https://doi.org/10.48550/arXiv.2507.18071.

A Extended related work

Table 2 summarizes the statistics and update rules. We expand these comparisons, then discuss token weighting, soft updates, and entropy analyses.

Policy objectives and ratio aggregation. Dr.GRPO (Liu et al., 2025) and REINFORCE++ (Hu et al., 2025) modify normalization and advantage estimation within policy gradient training. GPG (Chu et al., 2025) uses a direct policy gradient estimator without ratio clipping. These methods change the objective or advantage estimates; CPPO changes which token updates are retained.

GSPO (Zheng et al., 2025) clips a length-normalized sequence ratio, and GMPO (Zhao et al., 2026) uses geometric aggregation. MinPRO (Lei et al., 2026) replaces the product of preceding ratios with their minimum, then clips its product with the current ratio to weight the token's gradient. The minimum records an extreme sampled-token ratio within the prefix. CPPO instead sums weighted conditional divergences, so repeated deviations can reduce the allowance even when no single position has an extreme score.

Divergence constraints and update rules. Classical trust region and relative entropy methods (Kakade & Langford, 2002; Schulman et al., 2015; Achiam et al., 2017; Peters et al., 2010; Abdolmaleki et al., 2018; Song et al., 2019) constrain or regularize distributional change. Divergence information also guides clipping thresholds and rollback rules (Wang et al., 2019b;a). Empirical studies show that the resulting behavior depends on the optimization and implementation choices (Engstrom et al., 2020; Andrychowicz et al., 2021).

Table 2: Statistics and update rules in trust region policy optimization. Token divergence compares the rollout and current next-token distributions at the same prefix. N denotes the response length.

Method	Statistic	Update rule
PPO/GRPO	sampled-token ratio ρ_t	clip the ratio in the token surrogate
MinPRO	current ratio and minimum preceding ratio	clip their product to weight each token's gradient
DPPO	token divergence score $\widehat{D}_t$	retain tokens using a divergence threshold shared across positions
TROLL	KL on sparse next-token distributions	project distributions and use the projected probabilities in the objective
TRM-Max	response maximum $\max_{t\le N} D_t^{\mathrm{KL}}$	mask the response when its maximum divergence exceeds the threshold
TRM-Avg	response mean $N^{-1}\sum_{t=1}^{N} D_t^{\mathrm{KL}}$	mask the response when its mean divergence exceeds the threshold
CPPO (ours)	weighted score $w_t\widehat{D}_t$ and prefix sum S_t	retain tokens using position-weighted divergence and a cumulative prefix budget

For LLMs, DPPO (Qi et al., 2026) applies a token mask using Binary or Top-K approximations to TV or KL. CPPO changes the allowance for this kind of divergence score using position weights and a cumulative prefix budget. TRM (Li et al., 2025) uses the maximum or mean divergence over a response to decide whether to retain that response. These summary statistics do not retain the order of the divergence values; CPPO uses their accumulation along ordered prefixes. TROLL (Becker et al., 2026) changes the probabilities used in the objective through a differentiable KL projection on sparse next-token distributions. This projection acts on distributions, while the masks in DPPO, TRM, and CPPO select sampled updates.

Under exact TV constraints, Appendix B.8 compares uniform token limits, a budget on the final prefix, and a budget on the full response. It explains why controlling every prefix gives information beyond a single response aggregate.

Token weighting and soft updates. NFPO (Yoon et al., 2026) augments each token ratio with a product of ratios from subsequent tokens. FIPO (Ma et al., 2026) derives token weights from discounted future log-probability changes. Both use information from later positions to change the update weight. The cumulative constraint in CPPO uses preceding conditional divergence to determine the current token's allowance.

SAPO (Gao et al., 2025) replaces hard clipping with smooth attenuation based on the token ratio. Our soft mask ablation instead attenuates updates according to the joint CPPO budget score (Figure 4; Appendix C.3). It tests the form of the update rule while retaining the divergence conditions.

Entropy and learning dynamics. Studies of high-entropy reasoning tokens (Wang et al., 2025) and entropy dynamics (Cui et al., 2025) examine exploration and policy concentration during RLVR. Entropy describes uncertainty within one policy, whereas divergence measures change between policies. CPPO selects updates using the latter, together with token position and the preceding prefix.

B Proofs and additional theoretical results

This appendix follows the argument in Section 4. Appendices B.1–B.3 derive the surrogate identity and the position-dependent residual bound used in Sections 4.1 and 4.2. Appendix B.4 then proves Theorem 1 by converting all prefix constraints into a bound on that residual.

The remaining results explain the theorem's design implications. Appendices B.5 and B.6 verify the linear schedule and the threshold comparison in Section 4.3; Appendix B.7 derives the correction for the initial allowance used in Section 4.4. Appendix B.8 compares sequence-level constraints. Appendix B.9 derives a bound linear in the horizon and compares the smaller of the linear and

quadratic bounds under each constraint set. Appendix B.10 retains the suffix bound in product form and states its Abel condition. The discussion after the main proof identifies which step is tight.

We use the setup of Section 4: a fixed horizon $T \geq 2$, policies with common support, and $|R(x, y)| \leq \xi$ for $\xi > 0$. The quantities D_t and u_t denote exact conditional TV and its expectation under the rollout prefix distribution, respectively. Weights, token limits, and budget rates are deterministic. Distributional constraints hold μ-almost surely unless a result explicitly assumes only their expectations.

B.1 Finite-horizon performance difference identity

Equation (10) separates the change in expected reward into a token-level surrogate and its residual. We prove this identity first so that the subsequent analysis can focus on the interaction between a token ratio and the suffix that follows it.

Lemma 2 (Performance difference identity)**.** *Under common support, let* $\rho_{a:b} := \prod_{j=a}^{b} \rho_j$ *and* $\rho_{T+1:T} := 1$. *Then*

$$\begin{aligned} J(\pi) - J(\mu) &= L'_\mu(\pi) - \Delta(\mu, \pi), \\ L'_\mu(\pi) &:= \mathbb{E}_\mu\left[R(x, y) \sum_{t=1}^{T} (\rho_t - 1)\right], \\ \Delta(\mu, \pi) &:= \mathbb{E}_\mu\left[R(x, y) \sum_{t=1}^{T} (\rho_t - 1)(1 - \rho_{t+1:T})\right]. \end{aligned} \tag{16}$$

Hence $J(\pi) - J(\mu) \geq L'_\mu(\pi) - |\Delta(\mu, \pi)|$.

Proof. By importance sampling,

$$J(\pi) - J(\mu) = \mathbb{E}_\mu[R(x, y)(\rho_{1:T} - 1)]\,.$$

The reverse telescoping identity

$$\rho_{1:T} - 1 = \sum_{t=1}^{T} (\rho_t - 1)\rho_{t+1:T}$$

then gives

$$\begin{aligned} J(\pi) - J(\mu) &= \mathbb{E}_\mu\left[R(x, y) \sum_{t=1}^{T} (\rho_t - 1)\rho_{t+1:T}\right] \\ &= \mathbb{E}_\mu\left[R(x, y) \sum_{t=1}^{T} (\rho_t - 1)\right] - \mathbb{E}_\mu\left[R(x, y) \sum_{t=1}^{T} (\rho_t - 1)(1 - \rho_{t+1:T})\right]. \end{aligned}$$

The two terms on the last line are exactly $L'_\mu(\pi)$ and $\Delta(\mu, \pi)$. The lower bound follows from $-\Delta(\mu, \pi) \geq -|\Delta(\mu, \pi)|$. □

B.2 Suffix divergence under token-level limits

The residual contains a likelihood ratio for the entire remaining suffix, whereas the constraints act on individual token distributions. The next lemma connects these two quantities. Let $P_\mu^{t+1:T}(\cdot \mid s_{t+1})$ and $P_\pi^{t+1:T}(\cdot \mid s_{t+1})$ denote the two policies' conditional laws of the remaining tokens.

Lemma 3 (Suffix TV bound)**.** *Suppose* $D_j \leq \ell_j$ *pathwise for all* $j > t$, *with* $0 \leq \ell_j \leq 1$. *Then, for* d_{t+1}^μ*-almost every* s_{t+1},

$$D_{\mathrm{TV}}(P_\mu^{t+1:T}(\cdot \mid s_{t+1}), P_\pi^{t+1:T}(\cdot \mid s_{t+1})) \leq 1 - \prod_{j=t+1}^{T} (1 - \ell_j) \leq \sum_{j=t+1}^{T} \ell_j. \tag{17}$$

Proof. Construct a stepwise maximal coupling of the two suffix processes. As long as the two sampled suffixes have not disagreed, they share the same state s_j. The one-step disagreement probability at that state is

$$D_{\mathrm{TV}}(\mu(\cdot \mid s_j), \pi(\cdot \mid s_j)) = D_j \le \ell_j.$$

Therefore the conditional probability of no disagreement over the whole suffix is at least

$$\prod_{j=t+1}^{T} (1 - \ell_j).$$

The coupling characterization of total variation gives

$$\begin{aligned} D_{\mathrm{TV}}\big(P_\mu^{t+1:T}(\cdot \mid s_{t+1}), P_\pi^{t+1:T}(\cdot \mid s_{t+1})\big) &\le \Pr\big(Y_{t+1:T}^{\mu} \neq Y_{t+1:T}^{\pi} \mid s_{t+1}\big) \\ &\le 1 - \prod_{j=t+1}^{T} (1 - \ell_j). \end{aligned}$$

The union bound yields

$$1 - \prod_{j=t+1}^{T} (1 - \ell_j) \le \sum_{j=t+1}^{T} \ell_j.$$

□

B.3 Position-dependent surrogate error bound

We now apply the suffix bound to the interaction term in Lemma 2. Keeping the expected divergence at position t explicit gives the first inequality of Equation (11).

Proposition 4 (Position-dependent surrogate error bound)**.** *If $D_t \le \ell_t$ pathwise for all t, then*

$$|\Delta(\mu, \pi)| \;\le\; 4\xi \sum_{t=1}^{T-1} u_t \sum_{j=t+1}^{T} \ell_j. \tag{18}$$

Proof. Lemma 2 and $|R(x, y)| \le \xi$ give

$$|\Delta(\mu, \pi)| \le \xi \sum_{t=1}^{T} \mathbb{E}_\mu[|\rho_t - 1|\,|1 - \rho_{t+1:T}|]\,.$$

Fix $t < T$ and condition on s_{t+1}. The future likelihood ratio $\rho_{t+1:T}$ is the Radon–Nikodym derivative of the suffix law under π with respect to the suffix law under μ, so

$$\begin{aligned} \mathbb{E}_\mu[|1 - \rho_{t+1:T}| \mid s_{t+1}] &= \sum_{y_{t+1:T}} P_\mu^{t+1:T}(y_{t+1:T} \mid s_{t+1}) \left|1 - \frac{P_\pi^{t+1:T}(y_{t+1:T} \mid s_{t+1})}{P_\mu^{t+1:T}(y_{t+1:T} \mid s_{t+1})}\right| \\ &= 2D_{\mathrm{TV}}\big(P_\mu^{t+1:T}(\cdot \mid s_{t+1}), P_\pi^{t+1:T}(\cdot \mid s_{t+1})\big) \\ &\le 2 \sum_{j=t+1}^{T} \ell_j, \end{aligned}$$

The last line follows from Lemma 3; if a limit exceeds one, we first replace it by $\min\{\ell_j, 1\}$, which can only reduce the final sum. Conditional on s_{t+1}, the factor $|\rho_t - 1|$ is fixed, so the suffix bound can be applied inside the expectation. The remaining token factor satisfies

$$\begin{aligned} \mathbb{E}_{y_t \sim \mu(\cdot|s_t)}[|\rho_t - 1| \mid s_t] &= \sum_{y_t} \mu(y_t \mid s_t) \left|\frac{\pi(y_t \mid s_t)}{\mu(y_t \mid s_t)} - 1\right| \\ &= 2D_t. \end{aligned}$$

Taking expectation over $s_t \sim d_t^\mu$ gives $\mathbb{E}_\mu|\rho_t - 1| = 2u_t$. Combining the two displayed bounds,

$$\mathbb{E}_\mu[|\rho_t - 1|\,|1 - \rho_{t+1:T}|] \le 4u_t \sum_{j=t+1}^{T} \ell_j.$$

The term $t = T$ is zero because $\rho_{T+1:T} = 1$. Summing over $t = 1, \ldots, T-1$ proves the claim. □

Replacing each suffix limit by $\bar{\ell}$ gives the coefficient $\lambda_t = 4\xi\bar{\ell}(T-t)$ in Equation (11). The factor corresponding to the remaining horizon therefore enters through the suffix bound. These coefficients describe the position dependence of the residual estimate and motivate Equation (4).

B.4 Proof of the bound under prefix constraints

The residual bound still depends on the unknown divergences u_t. To obtain Theorem 1, we use the prefix budgets to control their weighted sum. Abel summation is useful here because it expresses this sum directly in terms of cumulative quantities, matching the form of the constraints.

Proof of Theorem 1. Step 1: bound the suffix contribution. The constraint $w_t D_t \le c_t$ in (12) implies

$$D_t \le \ell_t := \frac{c_t}{w_t}, \qquad \bar{\ell} := \max_t \ell_t.$$

By Proposition 4,

$$|\Delta(\mu,\pi)| \le 4\xi \sum_{t=1}^{T-1} u_t \sum_{j=t+1}^{T} \ell_j \le \sum_{t=1}^{T-1} \lambda_t u_t, \qquad \lambda_t := 4\xi\bar{\ell}(T-t).$$

Step 2: take expectations of the prefix constraints. Taking expectations in the prefix budget constraints gives, for every $m = 1, \dots, T-1$,

$$\sum_{j=1}^{m} w_j u_j \le \delta_b \sum_{j=1}^{m} w_j.$$

Define the centered prefix slack

$$A_m := \sum_{j=1}^{m} w_j(u_j - \delta_b), \qquad A_0 := 0.$$

Then $A_m \le 0$ for every $m < T$. Also define

$$\Delta A_t := A_t - A_{t-1} = w_t(u_t - \delta_b), \qquad r_t := \frac{\lambda_t}{w_t}.$$

Step 3: apply Abel summation. Using the previous definitions,

$$\begin{aligned}
\sum_{t=1}^{T-1} \lambda_t u_t - \delta_b \sum_{t=1}^{T-1} \lambda_t &= \sum_{t=1}^{T-1} \lambda_t (u_t - \delta_b) \\
&= \sum_{t=1}^{T-1} r_t \, \Delta A_t \\
&= r_{T-1} A_{T-1} + \sum_{t=1}^{T-2} (r_t - r_{t+1}) A_t.
\end{aligned}$$

By assumption, r_t is non-increasing, so $r_t - r_{t+1} \ge 0$; also $r_{T-1} \ge 0$ and $A_t \le 0$. Hence the right-hand side is non-positive, and therefore

$$\sum_{t=1}^{T-1} \lambda_t u_t \le \delta_b \sum_{t=1}^{T-1} \lambda_t.$$

Finally,

$$\sum_{t=1}^{T-1} \lambda_t = 4\xi\bar{\ell} \sum_{t=1}^{T-1} (T-t) = 2\xi T(T-1)\bar{\ell}.$$

Combining this residual bound with Lemma 2 gives

$$J(\pi) - J(\mu) \ge L'_\mu(\pi) - 2\xi T(T-1)\bar{\ell}\delta_b.$$

If $c_t \equiv \delta$ and $w_t \in [w_{\min}, 1]$, then

$$\bar{\ell} = \max_t \frac{\delta}{w_t} \leq \frac{\delta}{w_{\min}}.$$

Substituting this into the bound gives Equation (14). □

Sharpness of the weighted sum step. Among nonnegative sequences satisfying the expected prefix inequalities, $u_t = \delta_b$ makes every centered prefix sum A_m zero. It therefore attains the weighted sum bound in Step 3:

$$\sum_{t<T} \lambda_t u_t = \delta_b \sum_{t<T} \lambda_t.$$

This identifies where the full family of prefix constraints enters the comparison in Section 4.3. Attaining the complete surrogate error bound would also require equality in the reward and suffix divergence bounds and compatibility with the token limits. The sharpness statement here concerns the weighted sum step.

B.5 Abel condition for linear position weights

Section 4.3 applies the theorem to the linear schedule introduced in Section 3.1. The ratio of the residual coefficient to the position weight must be non-increasing. The following calculation verifies that condition for every allowed weight floor.

Proposition 5 (Abel condition for linear position weights)**.** *Let $T \geq 2$ and $0 < w_{\min} \leq 1$. For the linear schedule in Equation* (4)*, define $g_t = (T-t)/w_t$. Then*

$$g_t - g_{t+1} = \frac{w_{\min}}{w_t w_{t+1}} > 0, \qquad t = 1, \ldots, T-1.$$

Thus $r_t = 4\xi\bar{\ell}\, g_t$ is non-increasing and satisfies the monotonicity assumption of Theorem 1.

Proof. Write

$$\kappa := \frac{1 - w_{\min}}{T-1}, \qquad w_t = 1 - \kappa(t-1), \qquad g_t = \frac{T-t}{w_t}.$$

Then

$$\begin{aligned} g_t - g_{t+1} &= \frac{(T-t)w_{t+1} - (T-t-1)w_t}{w_t w_{t+1}} \\ &= \frac{w_t - \kappa(T-t)}{w_t w_{t+1}} \\ &= \frac{1 - \kappa(T-1)}{w_t w_{t+1}} \\ &= \frac{w_{\min}}{w_t w_{t+1}} > 0. \end{aligned}$$

Thus g_t is strictly decreasing. Multiplication by the nonnegative constant $4\xi\bar{\ell}$ gives the required monotonicity of r_t. □

B.6 Comparison with uniform token limits

The next two corollaries compare residual bounds under two ways of setting token limits. Both retain the prefix constraints and monotonicity assumption of Theorem 1. The comparison baseline has a uniform exact TV limit $D_t \leq \delta$ and no prefix budget. The first corollary holds that unweighted limit fixed; the second uses the constant weighted limit corresponding to the position component in Equation (7).

Corollary 6 (Uniform token-level threshold)**.** *Under the assumptions of Theorem 1, let $c_t = w_t\delta$, equivalently $\ell_t \equiv \delta$. Then $\bar{\ell} = \delta$ and*

$$J(\pi) - J(\mu) \;\geq\; L'_\mu(\pi) - 2\xi T(T-1)\delta\delta_b. \tag{19}$$

The residual constants are

$$C_{\mathrm{CPPO}} := 2\xi T(T-1)\delta\delta_b, \qquad C_{\mathrm{uniform}} := 2\xi T(T-1)\delta^2.$$

Thus

$$\frac{C_{\mathrm{CPPO}}}{C_{\mathrm{uniform}}} = \frac{\delta_b}{\delta}.$$

The residual bound under prefix constraints is smaller when $\delta_b < \delta \le 1$, with the same unweighted token limit in both cases.

Corollary 7 (Position-dependent token-level threshold)**.** *Under the assumptions of Theorem 1, let $c_t = \delta$, so that $\ell_t = \delta/w_t$. If $w_t \in [w_{\min}, 1]$, then $\bar{\ell} \le \delta/w_{\min}$ and*

$$J(\pi) - J(\mu) \;\ge\; L'_\mu(\pi) - 2\xi T(T-1)\frac{\delta}{w_{\min}}\delta_b. \tag{20}$$

Against the position-independent token-level divergence constant $C_{\mathrm{uniform}}(\delta) = 2\xi T(T-1)\delta^2$, the position-dependent bound uses

$$C_{\mathrm{impl}} := 2\xi T(T-1)\frac{\delta}{w_{\min}}\delta_b, \qquad \frac{C_{\mathrm{impl}}}{C_{\mathrm{uniform}}(\delta)} = \frac{\delta_b}{\delta w_{\min}}.$$

It is tighter whenever $\delta_b < \delta\, w_{\min}$. The linear schedule $w_t = 1 - \frac{1-w_{\min}}{T-1}(t-1)$ used throughout our experiments satisfies the monotonicity assumption of Theorem 1 by Proposition 5.

B.7 Bounds with an initial allowance

For $T \ge 2$, the divergence test at the first token in Equation (7) has threshold δ. Here we study the corresponding constraints with exact TV in place of the computed score. This adds a constant initial slack to the clean prefix inequality used in Theorem 1. The result supplies Equation (15) in Section 4.4 and quantifies the cost of that initial allowance.

Proposition 8 (Bound with an initial allowance)**.** *Assume the weighted constraint gives $D_t \le \bar{\ell}$ and, in expectation,*

$$\sum_{j=1}^{m} w_j u_j \le \delta_b W_m + \eta \qquad (m = 1, \dots, T-1)$$

for some constant slack $\eta \ge 0$. Under the same monotonicity condition as Theorem 1,

$$|\Delta(\mu,\pi)| \le 2\xi T(T-1)\bar{\ell}\delta_b \;+\; \eta\,\frac{4\xi\bar{\ell}(T-1)}{w_1}. \tag{21}$$

For the linear schedule, $w_1 = 1$. If the exact TV counterpart of Equation (7) *holds at every prefix, we may take $\eta = \delta$. The additional residual term is then $O(T\bar{\ell}\delta)$.*

Proof. Define

$$A_m := \sum_{j=1}^{m} w_j(u_j - \delta_b), \qquad A_0 := 0.$$

The assumed slack gives $A_m \le \eta$ for every $m < T$. As in the proof of Theorem 1, set

$$\lambda_t := 4\xi\bar{\ell}(T-t), \qquad r_t := \frac{\lambda_t}{w_t}.$$

Abel summation gives

$$\sum_{t=1}^{T-1}\lambda_t u_t - \delta_b\sum_{t=1}^{T-1}\lambda_t = r_{T-1}A_{T-1} + \sum_{t=1}^{T-2}(r_t - r_{t+1})A_t.$$

Because r_t is non-increasing, all coefficients on the right are nonnegative. Therefore

$$\begin{aligned} r_{T-1}A_{T-1} + \sum_{t=1}^{T-2}(r_t - r_{t+1})A_t &\le \eta\left(r_{T-1} + \sum_{t=1}^{T-2}(r_t - r_{t+1})\right) \\ &= \eta r_1 \\ &= \eta\,\frac{4\xi\bar{\ell}(T-1)}{w_1}. \end{aligned}$$

Combining this with Proposition 4 proves the displayed residual bound.

To identify the slack, assume the exact TV counterpart of Equation (7) holds at every prefix. It implies

$$P_m \le \delta + \delta_b W_{m-1} \le \delta + \delta_b W_m.$$

Taking expectations gives the proposition's prefix slack condition with $\eta = \delta$. □

Relation to the training objective. The bounds concern the reward surrogate L'_μ and a deterministic budget rate δ_b. The training objective in Equation (9) uses a token mask and group-normalized advantages. For the Base models, the budget rate is calibrated separately for each response. The analysis isolates how deterministic divergence budgets control the reward-surrogate residual.

B.8 Comparison with sequence-level constraints

This comparison concerns the exact TV constraints used to bound the surrogate residual. It expands the distinction between sequence-level and prefix-level aggregation in Section 6, which also motivates the controls in Section 5.4. A maximum, an average over the final prefix, and an average over the full response aggregate divergence differently; the distinction matters even before choosing how to mask gradients.

Maximum token divergence. The condition $\max_{t\le T} D_t \le \delta$ is equivalent to a uniform token limit and gives $u_t \le \delta$ at every position. The resulting residual bound and the bound with an additional prefix budget are, respectively,

$$C_{\max} = 2\xi T(T-1)\delta^2, \qquad C_{\text{prefix}} = 2\xi T(T-1)\delta\delta_b.$$

The second expression follows from Corollary 6 under its assumptions. Their ratio is δ_b/δ. This constraint is the exact TV analogue of the maximum divergence statistic used by TRM-Max.

A budget on the final prefix. The residual involves positions $1, \dots, T-1$. Write $W = \sum_{t<T} w_t$ and $r_t = \lambda_t / w_t$. If only the budget at $m = T-1$ is retained, then

$$\sum_{t<T} \lambda_t u_t = \sum_{t<T} r_t w_t u_t \le \left(\max_{t<T} r_t\right) \sum_{t<T} w_t u_t \le \delta_b W \max_{t<T} r_t.$$

Define the two bound expressions

$$C_{\text{terminal}}(w) := \delta_b W \max_{t<T} r_t, \qquad C_{\text{prefix}}(w) := \delta_b \sum_{t<T} \lambda_t.$$

The prefix expression is valid under Theorem 1. Since a maximum is at least its weighted average, $C_{\text{terminal}}(w) \ge C_{\text{prefix}}(w)$, with equality when r_t is constant over these positions. For uniform weights, the expressions reduce to

$$C_{\text{terminal}}(\mathbf{1}) = 4\xi\bar{\ell}(T-1)^2\delta_b, \qquad C_{\text{prefix}}(\mathbf{1}) = 2\xi T(T-1)\bar{\ell}\delta_b,$$

and hence

$$\frac{C_{\text{terminal}}(\mathbf{1})}{C_{\text{prefix}}(\mathbf{1})} = 2 - \frac{2}{T}. \tag{22}$$

This ratio compares the displayed bounds when the final budget covers the first $T-1$ positions.

A budget on the full response. A response mean, as used by TRM-Avg, covers all T positions. For its exact TV analogue, the condition $\sum_{t=1}^{T} D_t \le T\delta_b$ implies $\sum_{t<T} u_t \le T\delta_b$. With the same token limit $\bar{\ell}$,

$$\sum_{t<T} \lambda_t u_t \le 4\xi\bar{\ell}(T-1)T\delta_b = 2C_{\text{prefix}}(\mathbf{1}).$$

The factor differs because this budget includes the final token, whose coefficient in the residual is zero. The comparison isolates the effect of divergence aggregation under exact TV constraints. Relating these constants to KL constraints requires a TV–KL inequality.

B.9 LINEAR BOUND AND COMPARISON WITH UNIFORM LIMITS

The suffix TV in the proof of Proposition 4 is at most one. This gives the average-divergence bound of Qi et al. (2026, Theorem 3.2). Keeping the zero contribution at $t = T$ explicit, we obtain

$$|\Delta(\mu, \pi)| \le 4\xi \sum_{t=1}^{T-1} u_t. \tag{23}$$

A weighted prefix budget controls this sum even when the position weights decrease. The following result follows directly from the positive weight floor and the budget on the final prefix.

Proposition 9 (Linear bound under a weighted budget)**.** *Let $w_t \in [w_{\min}, 1]$ with $w_{\min} > 0$. If*

$$\sum_{t=1}^{T-1} w_t u_t \le \delta_b W_{T-1} + \eta, \qquad \eta \ge 0,$$

then

$$|\Delta(\mu, \pi)| \le \frac{4\xi}{w_{\min}} \big(\delta_b W_{T-1} + \eta\big). \tag{24}$$

Consequently, $J(\pi) - J(\mu)$ is at least $L'_\mu(\pi)$ minus the right-hand side.

Proof. For $t < T$, the conditional suffix likelihood ratio satisfies

$$\mathbb{E}_\mu[|1 - \rho_{t+1:T}| \mid s_{t+1}] = 2D_{\mathrm{TV}}(P_\mu^{t+1:T}, P_\pi^{t+1:T}) \le 2,$$

where both suffix laws are conditional on s_{t+1}. As in Proposition 4, $\mathbb{E}_\mu|\rho_t - 1| = 2u_t$. The residual identity therefore gives Equation (23). Since $u_t \ge 0$,

$$\sum_{t=1}^{T-1} u_t \le \frac{1}{w_{\min}} \sum_{t=1}^{T-1} w_t u_t \le \frac{\delta_b W_{T-1} + \eta}{w_{\min}}.$$

Combining these inequalities proves the bound; the reward statement follows from Lemma 2. □

Comparison of the linear and quadratic bounds. For $0 < \delta \le 1$, uniform token limits $D_t \le \delta$ give the residual bound

$$C^{\min}_{\mathrm{uniform}} := \min\big\{2\xi T(T-1)\delta^2,\ 4\xi(T-1)\delta\big\}. \tag{25}$$

The linear expression is slightly smaller than $4\xi T\delta$ because the final token has no suffix contribution. Under the assumptions of Corollary 7, the prefix constraints imply Proposition 9 with $\eta = 0$. Combining the two bounds gives

$$C^{\min}_{\mathrm{CPPO}} := \min\left\{2\xi T(T-1)\frac{\delta\delta_b}{w_{\min}},\ \frac{4\xi\delta_b W_{T-1}}{w_{\min}}\right\}. \tag{26}$$

The ratio of the quadratic expressions is $\delta_b/(\delta w_{\min})$. The ratio of the linear expressions is

$$\frac{\delta_b W_{T-1}}{\delta w_{\min}(T-1)} \le \frac{\delta_b}{\delta w_{\min}},$$

since $W_{T-1} \le T - 1$. Both branches therefore satisfy

$$C^{\min}_{\mathrm{CPPO}} \le \frac{\delta_b}{\delta w_{\min}}\, C^{\min}_{\mathrm{uniform}}. \tag{27}$$

Thus $\delta_b < \delta w_{\min}$ improves the comparison even when the uniform bound uses its linear branch. For the linear schedule,

$$W_{T-1} = \frac{T + (T-2)w_{\min}}{2},$$

so Equation (24) grows linearly with T for fixed positive budget parameters and weight floor. This result applies directly to the weights in Equation (4).

Initial allowance. For the exact TV counterpart of Equation (6), the budget at $m = T-1$ gives

$$\sum_{t=1}^{T-1} w_t u_t \le \delta + \delta_b W_{T-2}, \qquad |\Delta(\mu,\pi)| \le \frac{4\xi}{w_{\min}}\big(\delta + \delta_b W_{T-2}\big).$$

The initial allowance adds at most $4\xi\delta/w_{\min}$ to the linear bound, independently of T. This expression is smaller than the uniform linear bound when $\delta+\delta_b W_{T-2} < (T-1)\delta w_{\min}$. The quadratic correction is given in Proposition 8; the smaller of the two corrected bounds may be used.

For the linear schedule with weighted token limit δ, the comparison can also include the initial allowance explicitly. Define

$$C^{\min}_{\mathrm{CPPO,init}} := \min\left\{ 2\xi T(T-1)\frac{\delta\delta_b}{w_{\min}} + 4\xi(T-1)\frac{\delta^2}{w_{\min}},\ \frac{4\xi}{w_{\min}}(\delta + \delta_b W_{T-2}) \right\}. \tag{28}$$

The quadratic expression divided by its uniform counterpart is $\delta_b/(\delta w_{\min}) + 2/(T w_{\min})$. Since $W_{T-2} \le T-1$ and $1/(T-1) \le 2/T$ for $T \ge 2$, the corresponding ratio for the linear expressions is at most the same quantity. Thus, under the assumptions of Corollary 7 with the prefix budget replaced by $P_m \le \delta + \delta_b W_{m-1}$,

$$|\Delta(\mu,\pi)| \le C^{\min}_{\mathrm{CPPO,init}} \le \left(\frac{\delta_b}{\delta w_{\min}} + \frac{2}{T w_{\min}}\right) C^{\min}_{\mathrm{uniform}}. \tag{29}$$

The comparison therefore remains favorable with the initial allowance whenever $\delta_b/(\delta w_{\min}) + 2/(T w_{\min}) < 1$.

B.10 Product-form suffix bound

The main bound in Equation (11) uses a sum of suffix token limits. Lemma 3 also provides a product bound that never exceeds one. Keeping these coefficients can sharpen the estimate further. For $0 \le \ell_j \le 1$, define

$$\beta_t := 1 - \prod_{j=t+1}^{T} (1-\ell_j), \qquad \lambda^\beta_t := 4\xi\beta_t.$$

Keeping this bound in the proof of Proposition 4 yields

$$|\Delta(\mu,\pi)| \le 4\xi \sum_{t=1}^{T-1} u_t\beta_t = \sum_{t=1}^{T-1} \lambda^\beta_t u_t. \tag{30}$$

Under $P_m \le \delta_b W_m$ for $m < T$, Abel summation applies if the ratios $r^\beta_t := \lambda^\beta_t / w_t$ are non-increasing. To see this, use the centered prefix sums $A_m = \sum_{j\le m} w_j(u_j - \delta_b) \le 0$ from the main proof:

$$\begin{aligned}\sum_{t<T} \lambda^\beta_t u_t - \delta_b \sum_{t<T} \lambda^\beta_t &= r^\beta_{T-1} A_{T-1} + \sum_{t=1}^{T-2} (r^\beta_t - r^\beta_{t+1}) A_t \\ &\le 0.\end{aligned}$$

The last step uses both nonnegative coefficient differences and nonpositive prefix sums. With a uniform token limit $\ell_j = \delta \in (0,1]$, this gives

$$\begin{aligned}|\Delta(\mu,\pi)| &\le 4\xi\delta_b \sum_{t=1}^{T-1} \left[1 - (1-\delta)^{T-t}\right] \\ &= 4\xi\delta_b \left[(T-1) - \frac{(1-\delta)\big(1-(1-\delta)^{T-1}\big)}{\delta}\right] \le 4\xi(T-1)\delta_b.\end{aligned}$$

With uniform token limits and no prefix budget, Equation (30) instead gives the same sum multiplied by $4\xi\delta$. The ratio of these product bound expressions is therefore δ_b/δ.

This sharper comparison assumes that λ^β_t / w_t is non-increasing, as holds for constant weights. The coefficients β_t can saturate near one for long suffixes, so this condition must be checked separately for decreasing weights. Proposition 9 supplies the linear bound for the implemented schedule without this extra condition.

Table 3: Per-model rollout, update, and trust region configuration. Batch size and Minibatch count prompts; n gives responses per prompt. Updates is the number of optimizer updates per rollout iteration, and T is the maximum response length. The token-level scale δ is shared by DPPO and CPPO, and $w_{\min}$ is the floor for position weights. Budget is the fixed rate δ_b for Qwen3-1.7B and the calibration minimum $\delta_b^{\min}$ for the Base models, whose rates lie in $[\delta_b^{\min}, 2\delta_b^{\min}]$.

Model	Batch size	n	Minibatch	Updates	T	δ	Budget	$w_{\min}$
Qwen3-1.7B	64	8	32	2	8 k	0.15	0.015	0.8
Qwen3-1.7B-Base	64	8	32	2	8 k	0.15	0.020	0.8
Qwen3-8B-Base	128	8	32	4	8 k	0.15	0.020	0.8
Qwen3-30B-A3B-Base	256	16	32	8	16 k	0.20	0.020	0.8

C Experimental and implementation details

This appendix expands the experimental setup in Section 5.1 and the implementation of Algorithm 1. We give the shared protocol, baseline controls, and CPPO details, followed by policy deviation diagnostics, detailed benchmark results, and training traces. Reasoning excerpts then illustrate individual mask decisions.

C.1 Training and evaluation

Training configuration. All runs use the `verl`-compatible GRPO/DAPO trainer. For hyperparameters shared across methods, we follow the DPPO configuration. In particular, DPPO adopts the Dr. GRPO (Liu et al., 2025) setting for advantage computation, which centers rewards within each prompt group without dividing by the group standard deviation. We use AdamW with a maximum learning rate of 1×10^{-6}, no entropy regularizer, and no KL penalty relative to a reference policy in the loss. Each training step in the figures is one rollout iteration, which samples a batch of prompts and n responses per prompt under μ. The batch is then divided into minibatches for optimizer updates, as specified in Table 3. Training responses are sampled from the actor's untruncated softmax. Validation uses temperature 0.7, top-$p = 0.95$, and 16 responses per prompt.

Reward definition. The DAPO-Math-17K dataset (Yu et al., 2025) provides integer-valued reference answers. We use string pattern matching to extract the answer following the last `Answer:` prefix in each response, normalize it, and compare it with the reference answer. A correct response receives reward 1; a missing or incorrect answer receives 0. We use an outcome-only reward setting, with no overlong-response penalty or additional format reward term.

Evaluation protocol. We average correctness over the 16 responses per problem and then over problems to obtain each benchmark's $\mathrm{Avg@16}$. Tables 1 and 4 select the highest unweighted AIME24/25/26 mean within $[0, T^{\mathrm{stop}}]$, the matched training-step window. All three benchmark scores in Table 4 come from this same checkpoint.

C.2 Baseline settings and divergence scores

Likelihood ratio baselines. The GRPO baseline uses the Clip-Higher recipe with clipping interval $[1-\epsilon_{\mathrm{low}}, 1+\epsilon_{\mathrm{high}}]$ and $(\epsilon_{\mathrm{low}}, \epsilon_{\mathrm{high}}) = (0.2, 0.28)$. We denote it as GRPO in the tables. CISPO clips the importance weights in the policy gradient, whereas MinPRO weights the ratio for the current token by the minimum preceding token ratio. These methods use likelihood ratios without a distributional divergence score. The remaining baseline hyperparameters follow the recommendations in their original papers.

Reduced TV scores. The main DPPO and CPPO runs share the Top-K score with $K = 20$ and the nominal scale δ in Table 3. Following the partition construction of Qi et al. (2026), let $\mathcal{K}_t$ contain the top-K tokens under the rollout policy together with the sampled token. All remaining tokens are

Algorithm 1 CPPO mask computation

Require: response length $N \le T$, horizon $T \ge 2$, ratios $\rho_{1:N}$, advantages $\hat{A}_{1:N}$, scores $\widehat{D}_{1:N}$, token threshold δ, budget rate δ_b, weight floor $w_{\min}$.
Ensure: token mask $M_{1:N}$.
1: Initialize $S_0 \leftarrow 0$ and $W_0 \leftarrow 0$.
2: **for** $t = 1, \dots, N$ **do**
3: $\quad w_t \leftarrow 1 - (1 - w_{\min})(t-1)/(T-1)$.
4: $\quad b_t \leftarrow \min\{\delta, \delta + \delta_b W_{t-1} - S_{t-1}\}$.
5: $\quad M_t \leftarrow \mathbf{1}[\hat{A}_t(\rho_t - 1) \le 0 \ \vee \ w_t \widehat{D}_t \le b_t]$.
6: $\quad S_t \leftarrow S_{t-1} + w_t \widehat{D}_t$; $W_t \leftarrow W_{t-1} + w_t$.
7: **return** $M_{1:N}$.

grouped into one category. The resulting score and the Binary TV alternative are

$$\widehat{D}_t^{\text{Top-}K} = \frac{1}{2} \sum_{a \in \mathcal{K}_t} |\pi(a \mid s_t) - \mu(a \mid s_t)| + \frac{1}{2} |\pi(\mathcal{K}_t^c \mid s_t) - \mu(\mathcal{K}_t^c \mid s_t)|,$$
$$\widehat{D}_t^{\text{Binary}} = |\pi(y_t \mid s_t) - \mu(y_t \mid s_t)|.$$

Both are lower bounds on full-vocabulary TV; Binary TV uses the partition consisting of the sampled token and its complement. We use Binary TV in the divergence ablation in Figure 5 (right) and for the mask illustrations in Appendix C.6. The matched Top-K comparison fixes the approximation while changing the allocation and effective thresholds.

Sequence-level KL masks. TRM-Max and TRM-Avg use the maximum and average conditional KL over the valid response tokens, respectively. For valid length N, their sequence masks are

$$D_t^{\text{KL}} = \text{KL}\big(\mu(\cdot \mid s_t) \,\|\, \pi(\cdot \mid s_t)\big),$$
$$M_{\text{Max}}(y) = \mathbf{1}\left[\max_{1 \le t \le N} D_t^{\text{KL}} \le \delta\right], \qquad M_{\text{Avg}}(y) = \mathbf{1}\left[\frac{1}{N} \sum_{t=1}^{N} D_t^{\text{KL}} \le \delta_{\text{avg}}\right].$$

Each mask applies to the entire response, with the corresponding threshold recommended in the original paper. The Max&Avg control requires both inequalities to hold. Appendix B.8 compares maximum and cumulative constraints through their exact TV analogues.

C.3 CPPO IMPLEMENTATION

Budget calibration. Qwen3-1.7B uses the fixed rate $\delta_b = 0.015$. For the Base models, we calibrate one rate per response using the 90th percentile of divergence scores over its valid tokens:

$$\delta_b^{\text{seq}} = \min\left(2\delta_b^{\min}, \max\left(\delta_b^{\min}, P_{90}(\widehat{D}_{1:N})\right)\right). \tag{31}$$

The default minimum is $\delta_b^{\min} = 0.02$. The settings labeled $\delta_b = 0.02$ and $\delta_b = 0.03$ in Figure 5 (left) use this same calibration rule with different minima. The calibrated rate is held fixed across positions during one mask evaluation. Its quantile uses the completed response, whereas the cumulative divergence S_{t-1} in the mask uses only preceding positions.

Mask computation. Algorithm 1 uses the valid length N and the fixed maximum response length T. The position schedule extends to $T = 1$ as

$$w_t = w_{\min} + (1 - w_{\min}) \frac{T - t}{\max\{T - 1, 1\}},$$

so $w_1 = w_{\min}$ when $T = 1$. The upstream response mask excludes padding from both cumulative sums. Shifting them right by one position gives the quantities for the preceding prefix in Equation (7). All valid scores enter the sums before surrogate terms are masked. The statistics are recomputed at each loss evaluation and are not carried across optimizer steps.

Gradient computation. We implement the update using a ratio-weighted log probability loss with a cap on the ratio weights. For a batch of B responses,

$$\widehat{\mathcal{L}}_{\log} = -\frac{1}{BT}\sum_{i=1}^{B}\sum_{t=1}^{N_i}\mathrm{sg}\Big(M_{i,t}\min\{\rho_{i,t}, c\}\hat{A}_{i,t}\Big)\log\pi(y_{i,t} \mid s_{i,t}),$$

where i indexes responses. We retain the default ratio cap $c = 20$. For $\rho_{i,t} \le c$, minimizing this loss gives the same token contribution as maximizing Equation (9), since $\nabla_\theta \rho_t = \rho_t \nabla_\theta \log \pi_\theta(y_t \mid s_t)$ for fixed μ. Larger ratio weights are capped at c.

Soft mask variant. The ablation in Figure 4 (right) keeps position weighting and the prefix budget and replaces binary rejection with continuous attenuation. To combine the two inequalities in Equation (7), define the normalized score

$$x_t^{\mathrm{CPPO}} := \max\left\{\frac{w_t \widehat{D}_t}{\delta},\ \frac{S_t}{\delta + \delta_b W_{t-1}}\right\}. \tag{32}$$

The score is at most one exactly when both budget inequalities hold. Above one, it measures the larger relative violation. We retain the same directional condition. For outward updates that exceed the budget, we scale the gradient contribution by the inverse score:

$$M_t^{\mathrm{soft}} = \begin{cases} 1, & \hat{A}_t(\rho_t - 1) \le 0 \text{ or } x_t^{\mathrm{CPPO}} \le 1, \\ 1/x_t^{\mathrm{CPPO}}, & \text{otherwise.} \end{cases} \tag{33}$$

We substitute this mask into Equation (9), holding it fixed during differentiation. This ablation compares hard rejection with gradual attenuation under the same divergence conditions.

C.4 Position and budget diagnostics

Position-dependent probability changes. Figure 6 reports an offline diagnostic from Qwen3-30B-A3B rollouts after 200 GRPO rollout iterations. At each token position, we measure the sampled-token probability change

$$|\pi_\theta(y_t \mid s_t) - \mu(y_t \mid s_t)|,$$

where π_θ is the post-update training policy and μ is the rollout policy. The mean is taken across samples at each position bin. We then plot the maximum over each group of 20 consecutive position bins. The middle panel uses the 90th percentile across samples at each position bin, followed by the same maximum aggregation. The right panel gives the threshold δ/w_t induced by the CPPO position component.

Budget rates and prefix budget activity. Figure 7 tracks budget calibration and masking by the prefix budget in the three Base-model CPPO runs from the main experiments, using the configurations in Table 3. The top row reports the mean calibrated rate from Equation (31); the bottom row reports the fraction of all CPPO-masked tokens attributed to the prefix budget. The rates remain below the token-level scales ($\delta = 0.15$ or 0.20), with different trajectories across models. This prefix-budget share is highest early in these runs and becomes small later.

C.5 Detailed benchmark results and training traces

Training traces and benchmark scores. Figures 8–11 show the AIME24, AIME25, and AIME26 validation curves above the training reward, response length, and relative log probability error. The relative error is $\|\boldsymbol{\ell}^{\mathrm{train}} - \boldsymbol{\ell}^{\mathrm{rollout}}\|_2/(\|\boldsymbol{\ell}^{\mathrm{rollout}}\|_2 + 10^{-12})$, where the vectors contain training- and rollout-engine log probabilities for valid response tokens in a microbatch. Each training step reports the mean across microbatches, weighted by their response counts. Table 4 gives scores at the checkpoint selected by the protocol in Appendix C.1.

Per-benchmark rankings. The per-benchmark rankings vary (Table 4). On Qwen3-1.7B, CPPO leads on AIME24 and AIME26, while MinPRO leads on AIME25. On Qwen3-1.7B-Base, CPPO has the highest AIME25 and AIME26 scores but trails MinPRO and CISPO on AIME24. On Qwen3-8B-Base, CPPO leads on AIME25; MinPRO and CISPO lead on AIME24 and AIME26, respectively. For Qwen3-30B-A3B-Base, CPPO has the highest score on all three benchmarks. The aggregate ranking is therefore more consistent across model settings.

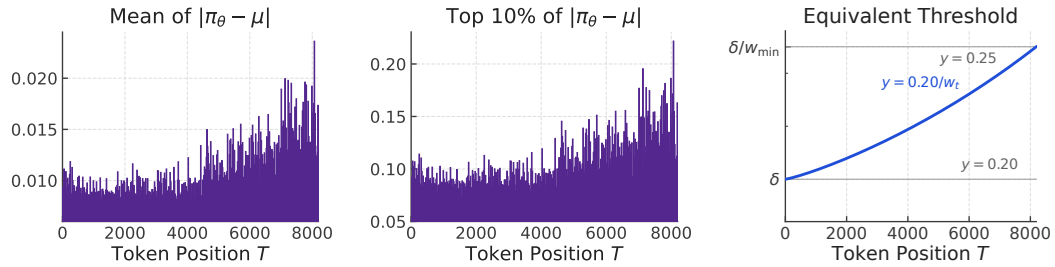

Figure 6: Position-conditioned policy deviation and the position-dependent token-level threshold. **Left**: mean absolute probability deviation across samples, followed by a maximum over 20 consecutive position bins. **Middle**: the 90th percentile of absolute probability deviation across samples, with the same aggregation over position bins. **Right**: the equivalent token-level threshold δ/w_t induced by the decreasing w_t schedule.

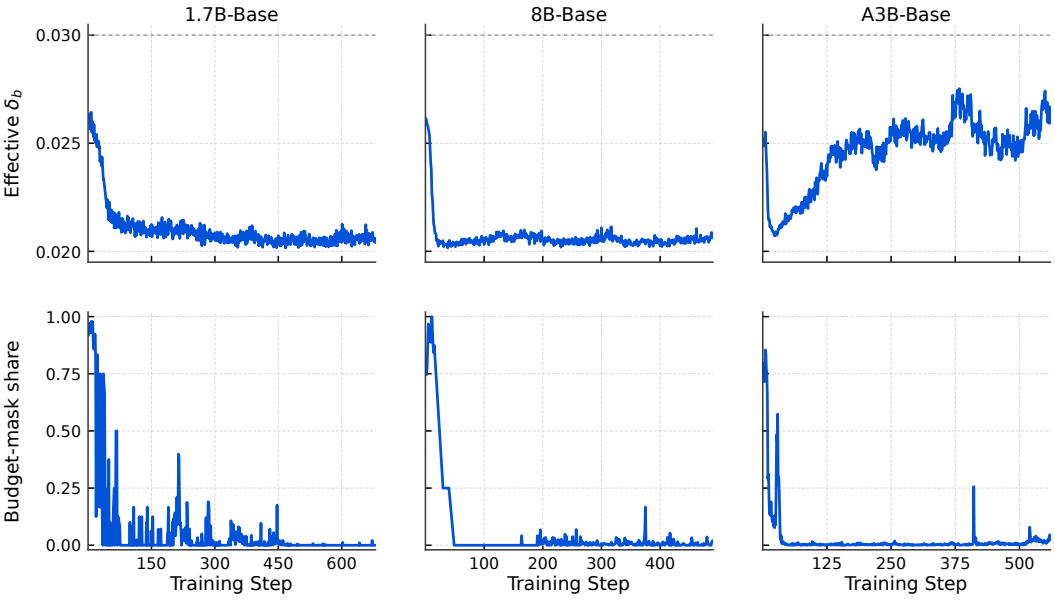

Figure 7: Adaptive budget diagnostics for the three Base-model CPPO runs in the main experiments. Each column is one model. Top: mean calibrated budget rate. Bottom: fraction of all CPPO-masked tokens attributed to the prefix budget.

Table 4: Per-benchmark validation scores (%, Avg@16). **AVG** is the best AIME24/25/26 mean within the matched training-step window; all benchmark scores come from that checkpoint. **Bold** and underlined values mark the best and second-best scores in each column and model block. *collapse* denotes CISPO's training divergence at step 215 on Qwen3-30B-A3B-Base.

Method	AIME24	AIME25	AIME26	AVG (↑)
Qwen3-1.7B ($T^{\text{stop}} = 1120$)				
GRPO	30.41	28.75	24.58	27.91
MinPRO	26.87	**32.71**	23.54	27.71
CISPO	30.62	28.12	27.71	28.82
DPPO	31.04	28.22	25.31	28.19
TRM-Max	26.46	22.91	26.25	25.21
TRM-Avg	29.79	25.41	25.41	26.87
CPPO (ours)	**34.79**	30.63	**30.21**	**31.88**
Qwen3-1.7B-Base ($T^{\text{stop}} = 680$)				
GRPO	11.25	6.87	8.54	8.89
MinPRO	**16.04**	10.41	6.66	11.04
CISPO	15.62	12.08	7.91	11.87
DPPO	13.96	8.33	10.41	10.90
TRM-Max	14.16	7.29	7.71	9.72
TRM-Avg	15.41	10.21	9.48	11.70
CPPO (ours)	15.21	**12.71**	**10.42**	**12.78**
Qwen3-8B-Base ($T^{\text{stop}} = 490$)				
GRPO	25.21	23.75	22.91	23.96
MinPRO	**35.21**	26.04	27.91	29.72
CISPO	32.91	25.41	**30.41**	29.58
DPPO	30.21	26.87	29.58	28.89
TRM-Max	28.33	24.37	27.50	26.73
TRM-Avg	31.46	25.00	27.50	27.98
CPPO (ours)	34.58	**30.00**	28.75	**31.11**
Qwen3-30B-A3B-Base ($T^{\text{stop}} = 560$)				
GRPO	43.75	32.71	38.12	38.19
MinPRO	52.91	38.54	52.91	48.12
CISPO	*collapse*	*collapse*	*collapse*	*collapse*
DPPO	57.29	39.58	50.83	49.23
TRM-Max	26.46	20.41	13.96	20.27
TRM-Avg	54.37	37.08	55.41	48.96
CPPO (ours)	**64.79**	**43.13**	**56.46**	**54.79**

C.6 Mask decisions in reasoning excerpts

Figures 12 and 13 use responses generated by Qwen3-1.7B in reasoning mode, with $\delta = 0.15$ and fixed $\delta_b = 0.015$. Figure 14 uses Qwen3-8B-Base with $\delta = 0.15$ and $\delta_b = 0.02$.

All three use Binary TV, $w_{\min} = 0.8$, and $T = 8192$. All three responses have positive sequence advantages and correct final answers.

Each box contains one token with its Binary TV below; blue and orange fills indicate retained and masked gradient contributions. Underscores denote spaces. Both rules retain gradients from tokens with $\rho_t \leq 1$ in these responses with positive advantages.

The first two excerpts contain incorrect reasoning steps. In Figure 12, the response claims that i^k is real only when k is a multiple of four, then corrects this to any even k. The selected token `a` has $\widehat{D}_t = 0.0680 < \delta$, but its weighted score exceeds the remaining prefix allowance: $w_t \widehat{D}_t \approx 0.0674 > b_t \approx 0.0574$. In Figure 13, the response mistakes adjacent sides AB and AD for opposite sides, inferring $AB = 1$ from $AD = 1$ before reconsidering. For the selected `must` token,

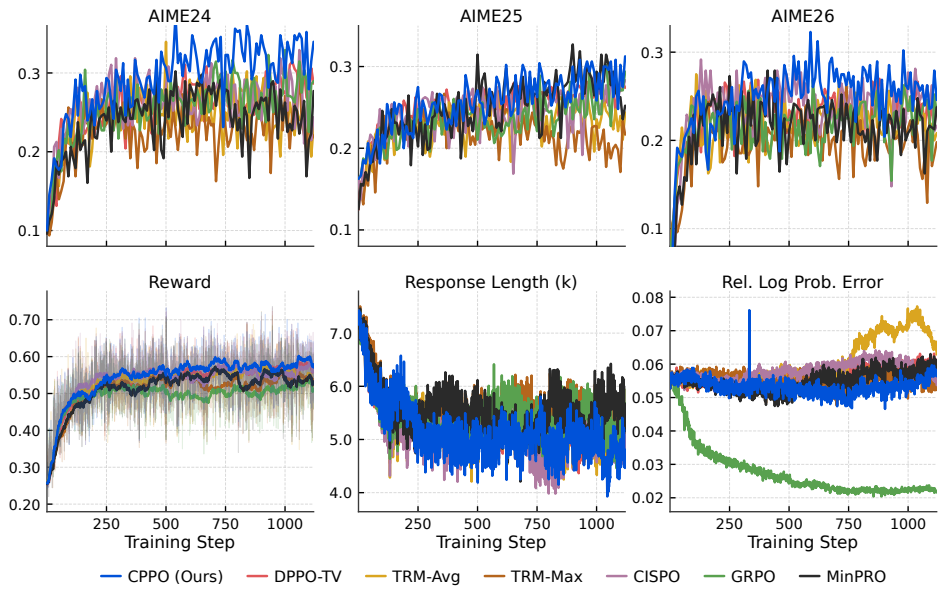

Figure 8: Complete training diagnostics for Qwen3-1.7B (post-trained). Top: AIME24, AIME25, and AIME26 validation Avg@16. Bottom: training reward, response length, and relative log probability error. Only reward is smoothed for readability.

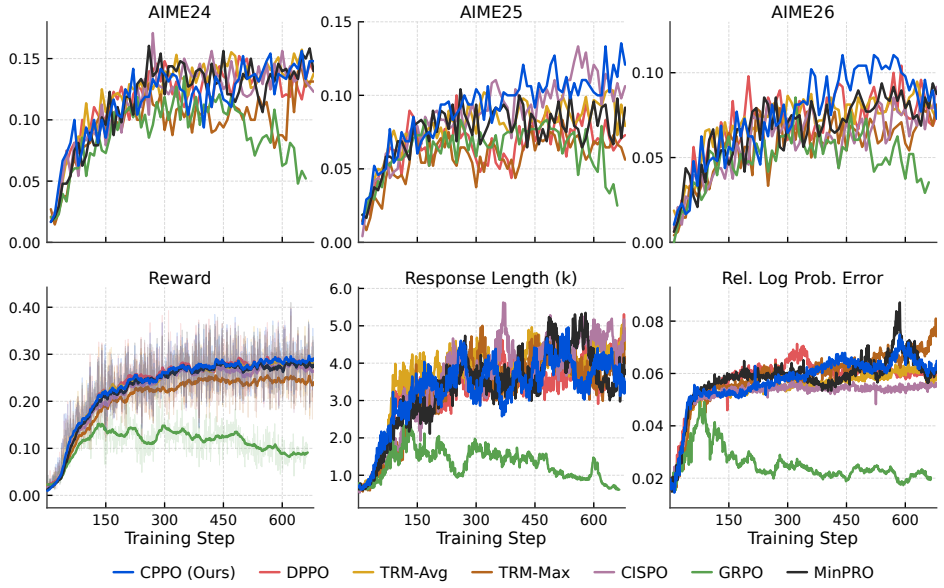

Figure 9: Complete training diagnostics for Qwen3-1.7B-Base. Top: AIME24, AIME25, and AIME26 validation Avg@16. Bottom: training reward, response length, and relative log probability error. Only reward is smoothed for readability.

$\widehat{D}_t = 0.0560 < \delta$, but $w_t \widehat{D}_t \approx 0.0559 > b_t \approx 0.0459$. Both selected tokens satisfy $\hat{A}_t(\rho_t - 1) > 0$. Uniform retains their gradient contributions, whereas CPPO masks them under the prefix budget test.

In Figure 14, the constant 2 occurs in a correct elimination step. Its divergence exceeds the uniform threshold but satisfies the position-weighted condition, so CPPO retains the token that Uniform masks. Correctness annotations describe the reasoning context; both masks use only policy statistics and advantages.

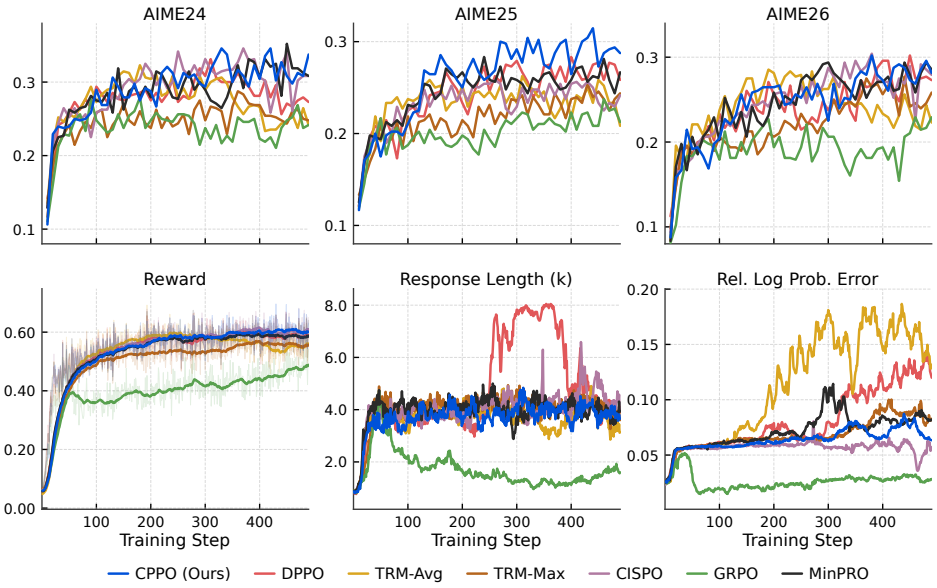

Figure 10: Complete training diagnostics for Qwen3-8B-Base. Top: AIME24, AIME25, and AIME26 validation Avg@16. Bottom: training reward, response length, and relative log probability error. Only reward is smoothed for readability.

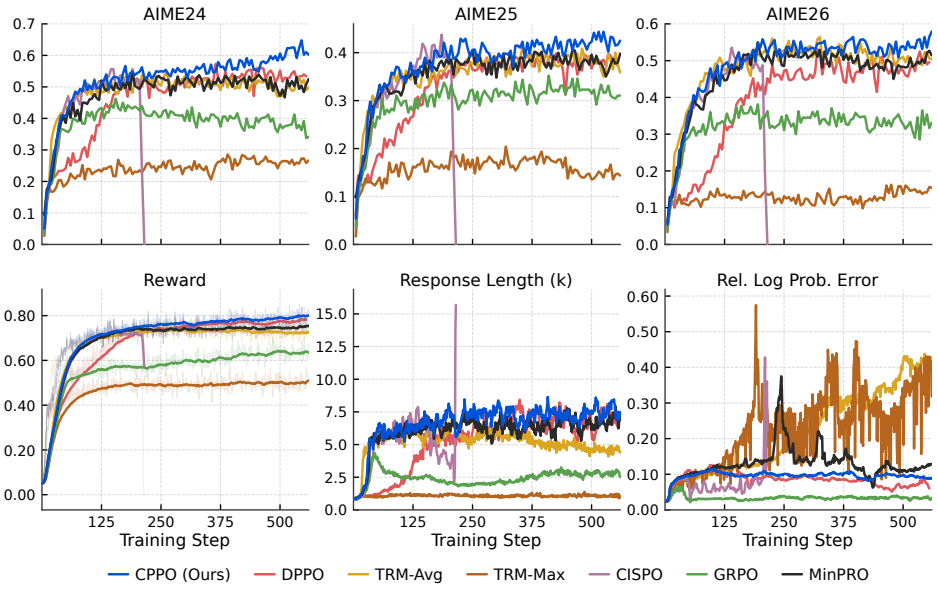

Figure 11: Complete training diagnostics for Qwen3-30B-A3B-Base. Top: AIME24, AIME25, and AIME26 validation Avg@16. Bottom: training reward, response length, and relative log probability error. Only reward is smoothed for readability.

Problem. Let S be the sum of all the real coefficients of the expansion of $(1 + ix)^{2009}$. What is $\log_2(S)$?

Response (abridged). Okay, let's try to solve this problem. …
Let me recall that $i^0 = 1$, $i^1 = i$, $i^2 = -1$, $i^3 = -i$, $i^4 = 1$, and then it repeats every four. So, the value of i^k is real **only** when k is **a** multiple of 4. ← an even k is enough. "a": direct error after preceding deviations.
Wait, no, actually, if k is even, then i^k is real.
…
Answer: 1004 Correct **adv > 0** (+0.5401)



Figure 12: Masking under the prefix budget in a Qwen3-1.7B reasoning excerpt with $\delta = 0.15$ and $\delta_b = 0.015$. The claim that i^k is real only for multiples of four is incorrect; the response then corrects it to even k. Uniform retains the selected `a` token, while CPPO masks its gradient contribution because the remaining prefix allowance is exceeded.

Problem. In rectangle $ABCD$, $AD = 1$, P is on AB, and DB and DP trisect angle ADC. The perimeter of triangle BDP is $k + m\sqrt{n}/3$ for integers k, m, n. Find $k + m + n$.

Response (abridged). Okay, so I need to solve this geometry problem about a rectangle ABCD. Let me try to visualize it first.
In rectangle ABCD, AD is 1. So, AD is one of the sides, and since it's a rectangle, **opposite sides are equal**. **Therefore, AB must also be 1?** ← AB and AD are not opposite sides. must: direct error after preceding deviations.
Wait, no. Wait, AD is given as 1. …
…
Answer: 9 Correct **adv > 0** (+0.5401)

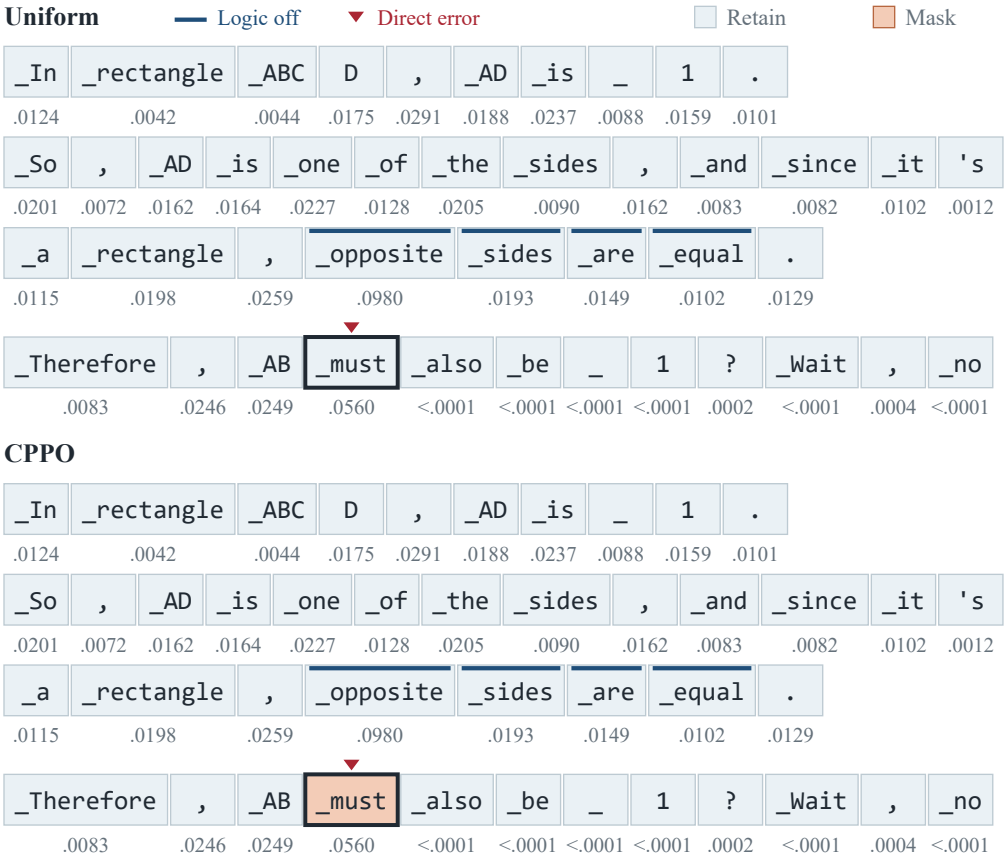

Figure 13: Masking under the prefix budget in a second Qwen3-1.7B reasoning excerpt, with $\delta = 0.15$ and $\delta_b = 0.015$. The response incorrectly infers $AB = 1$ from $AD = 1$ in a rectangle, although the two sides are adjacent. Uniform retains the selected `must` token; CPPO masks its gradient contribution because preceding divergence has reduced the prefix allowance.

Problem. Solve the system for a = 1 and sum the components of all solutions.
$x^2 + y^2 - 2z^2 = 2a^2$; $x + y + 2z = 4(a^2 + 1)$; $z^2 - xy = a^2$.

Response (abridged). Alright, I have this system of equations to solve:
…
$64 - 32z + 4z^2 - 2z^2 + 2 - 2z^2 = 2$
$64 - 32z + \underline{2} = 2$ ← highlighted 2 is correct
$66 - 32z = 2$
…
Answer: \$12\$ Correct **adv > 0** (+0.5401)



Figure 14: Position-weighted masking in a Qwen3-8B-Base excerpt with $\delta = 0.15$ and $\delta_b = 0.02$. The position-weighted allowance lets CPPO retain the correct constant 2 in $64 - 32z + 2 = 2$, which Uniform masks.