CONFORMAL PREDICTION-ENABLED SEQUENCE-LEVEL POLICY OPTIMIZATION FOR REASONING LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning improves the reasoning capability of large language models through verifiable rewards for mathematical and coding tasks. In token-level methods such as GRPO, clipping acts on individual token ratios while the reward and advantage apply to a complete response. Recent sequence-level methods align these units but still use fixed clipping bands, whose tail probabilities change as the score distribution evolves. We propose *Conformal Prediction-Enabled Sequence-level Policy Optimization* (CSPO), which adaptively calibrates sequence-level clipping bounds using recent training statistics. The resulting band targets a specified marginal tail budget and adapts to policy drift; calibration conditional on response length requires additional assumptions. We bound the probability that a fresh score falls outside the band by the target miscoverage level plus explicit CDF discrepancy and drift terms. Under additional moment conditions, this bound also controls the objective and gradient error relative to the unclipped GSPO power surrogate. Experiments with the Qwen3-1.7B series show improvements over GSPO on difficult mathematical reasoning benchmarks and ALFWorld, with clipping endpoints that adapt during training.

1 INTRODUCTION

Reinforcement learning from feedback has become a key component of large language model post-training. Early reinforcement learning from human feedback (RLHF) pipelines optimize policies against learned reward models derived from human preferences, enabling instruction-following and preference alignment beyond supervised fine-tuning (Christiano et al., 2017; Stiennon et al., 2020; Ouyang et al., 2022; Bai et al., 2022). More recently, reinforcement learning with verifiable rewards (RLVR) has emerged as an especially effective paradigm for reasoning tasks, where correctness can often be automatically checked by final-answer matching, symbolic verification, unit tests, or external graders. Such rewards have enabled large-scale reasoning-oriented post-training, as exemplified by recent reasoning models and open RL systems (Guo et al., 2025; Wen et al., 2025; Yu et al., 2025). In this setting, the reward is naturally defined at the *sequence level*: a complete solution is judged by whether the final answer and, in some cases, the reasoning trajectory satisfy the verification criterion.

Most RL post-training methods build on trust-region policy optimization. TRPO constrains the policy update through a KL trust region, while PPO replaces the hard constraint with a clipped surrogate objective that is simpler and more scalable (Schulman et al., 2015; 2017). Clipping limits the incentive to move probability ratios away from one, reducing the risk of overfitting sampled trajectories during repeated optimizer updates. For language models, GRPO simplifies PPO by removing the learned value function and estimating advantages through group-relative rewards, which substantially reduces memory and implementation complexity for long-context reasoning (Shao et al., 2024). However, GRPO inherits a token-level clipping design: each token ratio is clipped separately, while the advantage is typically a response-level scalar shared by all tokens in the completion. This creates a mismatch between the token-level control variable and the sequence-level reward signal.

A natural response to this mismatch is to optimize at the sequence level. GSPO constructs a response-level power multiplier from the likelihood ratio of the whole generated sequence and per-

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.

forms sequence-level clipping, rewarding, and optimization (Zheng et al., 2025). This aligns the unit of policy correction with the unit of reward and avoids treating individual tokens as independently rewarded decisions. Yet sequence-level clipping introduces a new challenge: the sequence log-ratio is a sum over tokens, so its scale depends strongly on the response length. FSPO identifies this length fairness issue and clips the sequence log-importance ratio with a band that scales as $\sqrt{L}$, where L is the output length (Mao et al., 2025). This is an important step toward fair sequence-level RL, but the clipping width remains determined by fixed hyperparameters. In long-chain-of-thought RL, the distribution of log-ratios can vary substantially across training stages as the policy improves, the entropy changes, and the model discovers new reasoning strategies. A fixed clipping band may be too conservative during exploration, suppressing rare but valuable responses, and too loose during unstable phases, allowing harmful updates.

We therefore calibrate the clipping band from recent policy shifts to target a specified marginal tail probability. This lets the numerical width change with the score distribution. Conformal prediction motivates this approach: empirical quantiles define intervals with controlled miscoverage under exchangeability, and adaptive conformal methods address calibration under distribution shift (Vovk et al., 2005; Shafer & Vovk, 2008; Angelopoulos & Bates, 2023; Gibbs & Candès, 2021). We use quantile calibration to select the sequence-level clipping band. Because training reuses data and changes the policy, our bounds explicitly retain CDF discrepancies and local drift. They are marginal over the response-length distribution; normalization by L alone does not imply calibration conditional on length.

We propose **C**onformal prediction-enabled **S**equence-level **P**olicy **O**ptimization (**CSPO**). Before each training step, CSPO uses a rolling buffer of retained training scores $W = \exp(S/L)$, where $S = \log[\pi_{\boldsymbol{\theta}}(\boldsymbol{y} \mid \boldsymbol{x})/\pi_{\boldsymbol{\theta}_{\text{old}}}(\boldsymbol{y} \mid \boldsymbol{x})]$ is the sequence-level log-ratio and $L = \text{len}(\boldsymbol{y})$. Given a target miscoverage level α, CSPO estimates lower and upper empirical quantiles $(q_{\text{lo}}, q_{\text{hi}})$ from completed blocks of retained scores, anchors the resulting interval at $W = 1$, and holds it fixed throughout the optimizer updates for that rollout batch. Scores computed on the training batch are retained for subsequent steps. The resulting sequence-level surrogate targets a marginal tail budget for fresh scores. Our contributions are threefold:

1. We identify the need for *adaptive* sequence-level clipping in RL post-training for reasoning language models, complementing prior work on sequence-level optimization and length-fair clipping.
2. We propose CSPO, which clips GSPO power scores using rolling empirical quantiles of retained training scores to target a specified marginal tail budget.
3. We bound tail probabilities for the mixture of scores within each training step in terms of α, CDF discrepancies, and local drift. Under additional moment conditions, these bounds control the objective and gradient error relative to the unclipped GSPO power surrogate. Empirically, CSPO improves Pass@1 over GSPO on difficult mathematical reasoning benchmarks and achieves the highest overall success rate on ALFWorld among the evaluated methods.

2 METHODOLOGY

A training step τ consists of one rollout batch and E_τ optimizer updates. The rollout parameter $\boldsymbol{\theta}_{\tau,0}$ remains fixed in every denominator; $\boldsymbol{\theta}_{\tau,k}$ is the current parameter before update k, for $k = 0, \ldots, E_\tau - 1$, and $\boldsymbol{\theta}_{\tau+1,0} = \boldsymbol{\theta}_{\tau,E_\tau}$. We use t for the current training step and $\tau < t$ for a retained step. At a fixed (t, k), the shorter notation $\boldsymbol{\theta} = \boldsymbol{\theta}_{t,k}$ and $\boldsymbol{\theta}_{\text{old}} = \boldsymbol{\theta}_{t,0}$ is used in the surrogate formulas. Let $F_{t,k}$ be the population CDF of the score evaluated at this policy pair, as defined in equation 9. Its band risk is

$$p_{t,k}(a, b) = F_{t,k}(a^-) + 1 - F_{t,k}(b). \tag{1}$$

We first define the sequence score, then the rolling buffer used to select future clipping bands. The same training data supply both the policy update and the retained scores.

Let $\boldsymbol{x} \sim p$ be a prompt and $\boldsymbol{y}_i = (y_{i,1}, \ldots, y_{i,L_i}) \sim \pi_{\boldsymbol{\theta}_{\text{old}}}(\cdot \mid \boldsymbol{x})$ one of G responses with sequence reward $\mathcal{R}_i$. We use the group-normalized advantage $\widehat{A}_i = [\mathcal{R}_i - \text{mean}_j(\mathcal{R}_j)]/[\text{std}_j(\mathcal{R}_j) + \epsilon_A]$ and token ratio $r_{i,\ell}(\boldsymbol{\theta}) = \pi_{\boldsymbol{\theta}}(y_{i,\ell} \mid \boldsymbol{h}_{i,\ell})/\pi_{\boldsymbol{\theta}_{\text{old}}}(y_{i,\ell} \mid \boldsymbol{h}_{i,\ell})$. Appendix B.1 records the full factorization

and token-level GRPO reference objective. Sequence-level optimization uses a single multiplier for each complete response. For response i, define the sequence log-ratio, the exact sequence likelihood ratio, and the GSPO power score by

$$S_i(\boldsymbol{\theta}) = \log \frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{y}_i \mid \boldsymbol{x}_i)}{\pi_{\boldsymbol{\theta}_{\text{old}}}(\boldsymbol{y}_i \mid \boldsymbol{x}_i)} = \sum_{\ell=1}^{L_i} \log r_{i,\ell}(\boldsymbol{\theta}),\ R_i(\boldsymbol{\theta}) = \exp(S_i(\boldsymbol{\theta}))\,,\ W_i(\boldsymbol{\theta}) = R_i(\boldsymbol{\theta})^{1/L_i}. \tag{2}$$

Here R_i is the exact density ratio between sequence distributions when absolute continuity holds. The GSPO multiplier W_i is the geometric mean of the token ratios and generally differs from an exact importance weight.

In indexed form, the stored score is

$$S_{\tau,k}(\boldsymbol{x}, \boldsymbol{y}) = \log \frac{\pi_{\boldsymbol{\theta}_{\tau,k}}(\boldsymbol{y} \mid \boldsymbol{x})}{\pi_{\boldsymbol{\theta}_{\tau,0}}(\boldsymbol{y} \mid \boldsymbol{x})}, \qquad W_{\tau,k}(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(\frac{S_{\tau,k}(\boldsymbol{x}, \boldsymbol{y})}{L(\boldsymbol{y})}\right). \tag{3}$$

This is the cached score for the length exponent $\beta = 1$ analyzed here. With a general exponent, the cached value is $\exp(S_{\tau,k}/L^{\beta})$. All score CDFs refer to this exponentiated, length-normalized value. Numerical capping inside the loss does not redefine the cached score.

With fixed constants $\epsilon_{\text{low}}, \epsilon_{\text{high}} > 0$, the response-level GSPO objective can be written as

$$J_{\text{seq}}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{x}, \{\boldsymbol{y}_i\} \sim \pi_{\boldsymbol{\theta}_{\text{old}}}} \left[\frac{1}{G} \sum_{i=1}^{G} \min\Big\{ W_i(\boldsymbol{\theta}) \widehat{A}_i,\ \text{clip}(W_i(\boldsymbol{\theta}), 1 - \epsilon_{\text{low}}, 1 + \epsilon_{\text{high}})\, \widehat{A}_i \Big\} \right]. \tag{4}$$

This fixed-band objective is the baseline that CSPO adapts. In particular, CSPO reduces exactly to this clipping rule whenever its calibrated endpoints are fixed at $1 - \epsilon_{\text{low}}$ and $1 + \epsilon_{\text{high}}$.

The per-update risk is $p_{t,k}^{\text{band}} := p_{t,k}(a_t, b_t)$. The step-level risk p_t^{band} will denote its mixture weighted by the numbers of scores actually retained at each update. It is a *potential* clipping risk: the actual constant branch also depends on the advantage sign. The parameter $\alpha = \alpha_{\text{lo}} + \alpha_{\text{hi}}$ specifies the nominal tail mass used to regularize the surrogate. It does not specify a KL radius or guarantee reward improvement.

The rollout batch $\mathcal{D}_t$ is generated once from $\pi_{\boldsymbol{\theta}_{t,0}}$ and may be reused across optimizer updates. Let $m_{\tau,k}$ be the number of score entries actually retained from forward evaluations at $\boldsymbol{\theta}_{\tau,k}$, and let $\mathcal{C}_\tau = \{k : m_{\tau,k} > 0\}$. A forward evaluation that is not collected contributes no entries. Repeatedly retained scores from the same response count as separate entries, each at the parameter used in that forward evaluation. Write

$$\mathcal{A}_{\tau,k} = (W_{\tau,k,j})_{j=1}^{m_{\tau,k}}, \qquad \mathcal{A}_\tau = \biguplus_{k \in \mathcal{C}_\tau} \mathcal{A}_{\tau,k}, \qquad m_\tau = \sum_{k \in \mathcal{C}_\tau} m_{\tau,k}.$$

Finalizing a block concatenates the recorded values; it does not recompute them at $\boldsymbol{\theta}_{\tau,E_\tau}$. Only completed nonempty blocks are included in the rolling buffer.

Let $\mathcal{I}_t \subseteq \{1, \ldots, t-1\}$ index the completed nonempty blocks retained before step t, and flatten each $\mathcal{A}_\tau$ as $(W_{\tau,j})_{j=1}^{m_\tau}$. Thus

$$\mathcal{A}_\tau = (W_{\tau,j})_{j=1}^{m_\tau}, \qquad \mathcal{B}_t = \biguplus_{\tau \in \mathcal{I}_t} \mathcal{A}_\tau, \qquad n_t = \sum_{\tau \in \mathcal{I}_t} m_\tau,$$

with empirical CDF $\widehat{F}_{\mathcal{B}_t}(w) = \frac{1}{n_t} \sum_{\tau \in \mathcal{I}_t} \sum_{j=1}^{m_\tau} \mathbf{1}\{W_{\tau,j} \le w\}$. We use the inverse empirical CDF to define the quantiles: $Q_p(\mathcal{B}_t) = \inf\{w : \widehat{F}_{\mathcal{B}_t}(w) \ge p\}, p \in (0,1)$. For $\alpha_{\text{lo}}, \alpha_{\text{hi}} > 0$ with $\alpha_{\text{lo}} + \alpha_{\text{hi}} = \alpha < 1$, the raw endpoints are

$$q_{\text{lo},t} = Q_{\alpha_{\text{lo}}}(\mathcal{B}_t), \qquad q_{\text{hi},t} = Q_{1-\alpha_{\text{hi}}}(\mathcal{B}_t). \tag{5}$$

A symmetric allocation uses $\alpha_{\text{lo}} = \alpha_{\text{hi}} = \alpha/2$; asymmetric allocations are allowed but must be fixed before selecting the clipping quantiles.

The unrestricted empirical quantiles above admit the blockwise result in Theorem B.2. To express an optional pooled certificate on a finite set of thresholds, grid mode predeclares a finite set $\mathcal{G}_M =$

$\{g_1 < \cdots < g_M\} \subset (0, \infty)$ before training. Define outward rounding by $\lfloor u \rfloor_{\mathcal{G}} = \max\big(\{0\} \cup \{g \in \mathcal{G}_M : g \le u\}\big)$, $\lceil u \rceil_{\mathcal{G}} = \min\big(\{+\infty\} \cup \{g \in \mathcal{G}_M : g \ge u\}\big)$, and set

$$\widetilde{q}_{\mathrm{lo},t} = \lfloor q_{\mathrm{lo},t} \rfloor_{\mathcal{G}}, \qquad \widetilde{q}_{\mathrm{hi},t} = \lceil q_{\mathrm{hi},t} \rceil_{\mathcal{G}}. \tag{6}$$

Outward rounding can only enlarge the interval. A logarithmically spaced grid containing 1 makes the approximation multiplicative; reaching 0 or $+\infty$ disables clipping on that side and must be reported. Because each rollout step begins at $W_{t,0} = 1$, the operational endpoints are

$$a_t = \min\{\widetilde{q}_{\mathrm{lo},t}, 1\}, \qquad b_t = \max\{\widetilde{q}_{\mathrm{hi},t}, 1\}. \tag{7}$$

When grid mode is disabled, set $\widetilde{q}_{\mathrm{lo},t} = q_{\mathrm{lo},t}$ and $\widetilde{q}_{\mathrm{hi},t} = q_{\mathrm{hi},t}$. If the buffer is empty, use warm-start bounds $(1 - c, 1 + c)$, $c \in (0, 1)$. The CSPO surrogate is

$$J_{\mathrm{cspo}}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{x}, \{\boldsymbol{y}_i\} \sim \pi_{\boldsymbol{\theta}_{t,0}}} \left[\frac{1}{G} \sum_{i=1}^{G} \min\Big\{ W_i(\boldsymbol{\theta}) \widehat{A}_i, \mathrm{clip}(W_i(\boldsymbol{\theta}), a_t, b_t) \widehat{A}_i \Big\} \right], \tag{8}$$

where (a_t, b_t) is fixed throughout the optimizer updates of step t. The empirical objective is evaluated at $\boldsymbol{\theta}_{t,k}$ on the responses used by update k. Algorithm 1 retains the scores selected by the existing collection schedule and appends their concatenation for subsequent steps.

Algorithm 1 CSPO with a rolling buffer of training scores and grid mode

Require: tail levels $\alpha_{\mathrm{lo}}, \alpha_{\mathrm{hi}}$, horizon K, predeclared grid $\mathcal{G}_M$, warm-start radius $c \in (0, 1)$

Initialize block buffer $\mathcal{B} \leftarrow \varnothing$
for training step $t = 1, 2, \ldots$ **do**
  Freeze the rollout parameter $\boldsymbol{\theta}_{\mathrm{old}} \leftarrow \boldsymbol{\theta}_{t,0}$
  If $\mathcal{B} = \varnothing$, use $(a_t, b_t) = (1 - c, 1 + c)$; otherwise compute raw quantiles by equation 5, round outward by equation 6, and anchor by equation 7
  Generate the rollout batch $\mathcal{D}_t$ from $\pi_{\boldsymbol{\theta}_{t,0}}$ and compute its group-relative advantages
  for optimizer update $k = 0, \ldots, E_t - 1$ **do**
    Evaluate the loss and scores $W_{t,k}$ on the responses used by this update, with numerator $\pi_{\boldsymbol{\theta}_{t,k}}$ and denominator $\pi_{\boldsymbol{\theta}_{t,0}}$
    If collection is enabled, retain the corresponding $m_{t,k}$ score entries in $\mathcal{A}_{t,k}$
    Apply the optimizer update to obtain $\boldsymbol{\theta}_{t,k+1}$
  end for
  Finalize $\mathcal{A}_t = \biguplus_{k \in \mathcal{C}_t} \mathcal{A}_{t,k}$ at its recorded values and set $\boldsymbol{\theta}_{t+1,0} \leftarrow \boldsymbol{\theta}_{t,E_t}$
  Record empirical diagnostics; append the nonempty $\mathcal{A}_t$ for subsequent steps and retain the most recent K blocks
end for

3 Theoretical Results

The analysis first bounds tail probabilities for the mixture of scores collected within a training step, accounting for CDF discrepancies and drift. Moment conditions then translate these bounds into objective and gradient error relative to the GSPO power surrogate. Appendix B gives the results for length groups, quantile tracking, exchangeable calibration, and error accumulation across updates under their additional assumptions.

Fix the realized optimization path. For each (τ, k), define the population CDF by pushing the rollout law through the score map in equation 3:

$$F_{\tau,k}(w) = \int \mathbf{1}\{W_{\tau,k}(\boldsymbol{x}, \boldsymbol{y}) \le w\} \, p(d\boldsymbol{x}) \pi_{\boldsymbol{\theta}_{\tau,0}}(d\boldsymbol{y} \mid \boldsymbol{x}). \tag{9}$$

The parameters in this integral are held at their realized values, including any dependence on $\mathcal{D}_\tau$. For $m_\tau > 0$, set $\nu_{\tau,k} = m_{\tau,k}/m_\tau$ and define

$$F_\tau = \sum_{k \in \mathcal{C}_\tau} \nu_{\tau,k} F_{\tau,k}, \qquad \widehat{F}_\tau = \sum_{k \in \mathcal{C}_\tau} \nu_{\tau,k} \widehat{F}_{\tau,k}, \qquad \widehat{F}_{\tau,k}(w) = \frac{1}{m_{\tau,k}} \sum_{j=1}^{m_{\tau,k}} \mathbf{1}\{W_{\tau,k,j} \le w\}. \tag{10}$$

Thus F_τ is the population mixture associated with the actual collection weights, not the score law of a single policy pair. The weights count entries rather than distinct responses. In particular, repeated scoring does not create additional independent rollouts, and the entries in $\mathcal{A}_\tau$ are not assumed i.i.d.

Let $\mathcal{H}_\tau$ denote the history after step τ, including its parameter path, collection counts, and stored scores. Conditional on this history, a population score $W_\tau^\star$ is defined by selecting k with probabilities $\nu_{\tau,k}$ and evaluating $W_{\tau,k}$ on an independent hypothetical pair $(\boldsymbol{x}_\tau^\star, \boldsymbol{y}_\tau^\star)$ from $p\pi_{\boldsymbol{\theta}_{\tau,0}}$, with $L_\tau^\star = L(\boldsymbol{y}_\tau^\star)$. This defines F_τ and its risk without an additional algorithmic sampling step:

$$p_\tau(a,b) = F_\tau(a^-) + 1 - F_\tau(b) = \sum_{k\in\mathcal{C}_\tau} \nu_{\tau,k} p_{\tau,k}(a,b), \qquad p_\tau^{\mathrm{band}} = p_\tau(a_\tau, b_\tau). \tag{11}$$

For the retained block weights $\lambda_{\tau,t} = m_\tau/n_t$, define the mixture CDF and its current-round discrepancy by

$$\overline{F}_t(w) = \sum_{\tau\in\mathcal{I}_t} \lambda_{\tau,t} F_\tau(w), \qquad D_t = \|\overline{F}_t - F_t\|_\infty. \tag{12}$$

The weighted drift D_t is the quantity needed by the quantile argument and obeys $D_t \le \Delta_K$ under Assumption 3.1; retaining it avoids replacing an average discrepancy by the worst block unnecessarily.

To separate the empirical quantile argument from dependence within a reused rollout batch, define the realized CDF discrepancies

$$\xi_{\tau,k} = \|\widehat{F}_{\tau,k} - F_{\tau,k}\|_\infty, \qquad \xi_\tau = \|\widehat{F}_\tau - F_\tau\|_\infty \le \sum_{k\in\mathcal{C}_\tau} \nu_{\tau,k}\xi_{\tau,k}. \tag{13}$$

These are quantities defined on the realized training path, not assumed sampling rates. The bounds below retain them explicitly; the number of cached entries alone does not bound them under adaptive reuse.

The following condition bounds drift between the step-level mixture CDFs.

Assumption 3.1. For a step t with $m_t > 0$, let $\mathcal{I}_t$ contain the latest $K_t = |\mathcal{I}_t|$ retained blocks. Almost surely along the training trajectory,

$$\sup_{w\in\mathbb{R}} |F_t(w) - F_\tau(w)| \le \Delta_K, \qquad \forall \tau \in \mathcal{I}_t, \tag{14}$$

where $\Delta_K \in [0,1]$ is the maximum local CDF drift over the retained window.

Theorem B.2 gives the pathwise, tie-robust bound $p_t^{\mathrm{band}} \le \alpha + 2\varepsilon_t + 2D_t$, where $\varepsilon_t = \sum_{\tau\in\mathcal{I}_t}(m_\tau/n_t)\xi_\tau$. The risk concerns the current step mixture F_t. Corollary B.7 quantifies conservatism through its largest endpoint atom. Theorem B.2 is marginal over response length; Proposition B.3 and Corollary B.8 state the additional discrepancy required for length-stratified transfer.

3.1 Optimization consequences of calibrated clipping

We next connect the calibrated tail event to the optimization objective actually used by CSPO. Fix an update (t,k), hold $\boldsymbol{\theta} = \boldsymbol{\theta}_{t,k}$ and $\boldsymbol{\theta}_{\mathrm{old}} = \boldsymbol{\theta}_{t,0}$ at their realized values, and fix finite thresholds $0 < a \le 1 \le b < \infty$. Rounds in which grid rounding returns $a = 0$ or $b = +\infty$ simply disable clipping on that side and are excluded from the finite-endpoint distortion statements below. Draw $\boldsymbol{x} \sim p(\boldsymbol{x}), \boldsymbol{y}_{1:G} \overset{\text{i.i.d.}}{\sim} \pi_{\boldsymbol{\theta}_{\mathrm{old}}}(\cdot \mid \boldsymbol{x}), U \sim \mathrm{Unif}\{1,\dots,G\}$, where U is conditionally independent of the response group. For each response, define $R_{\boldsymbol{\theta},i} = \frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{y}_i|\boldsymbol{x})}{\pi_{\boldsymbol{\theta}_{\mathrm{old}}}(\boldsymbol{y}_i|\boldsymbol{x})}, W_{\boldsymbol{\theta},i} = R_{\boldsymbol{\theta},i}^{1/L_i}$, compute the group-relative advantages $(\widehat{A}_1,\dots,\widehat{A}_G)$, and set

$$(\boldsymbol{y}, \widehat{A}, L, R_{\boldsymbol{\theta}}, W_{\boldsymbol{\theta}}) = (\boldsymbol{y}_U, \widehat{A}_U, L_U, R_{\boldsymbol{\theta},U}, W_{\boldsymbol{\theta},U}).$$

All expectations below are taken over this extended probability space; equivalently, they average $G^{-1}\sum_i$ within a random prompt group. The marginal CDF of this fresh score is $F_{t,k}$; the step-level F_t additionally averages over collected update indices. Assume $\pi_{\boldsymbol{\theta}}$ is absolutely continuous with respect to $\pi_{\boldsymbol{\theta}_{\mathrm{old}}}$ and $L \ge L_{\min} \ge 2$ almost surely. Here $R_{\boldsymbol{\theta}}$ is the likelihood ratio of the complete sampled object, including the declared EOS, termination, and truncation convention. The analysis

requires $L \geq 2$, which can be enforced by a documented effective-length rule; otherwise Lemma 3.2 must be replaced by an explicit moment assumption. Before bounding clipping distortion, we need a moment guarantee that does not itself assume a bounded likelihood ratio. The power $1/L$ supplies a uniform second-moment bound for the GSPO multiplier.

Lemma 3.2. *Under the preceding conditions,*

$$\mathbb{E}_{\boldsymbol{\theta}_{\mathrm{old}}}[W_{\boldsymbol{\theta}}^2] \leq 1 + \frac{2}{L_{\min}}. \tag{15}$$

To translate this moment control into objective fidelity, define

$$\ell_{a,b}(w, A) = \min\{wA, \mathrm{clip}(w, a, b)A\},\ \mathcal{J}_{\mathrm{G}}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}_{\mathrm{old}}}[W_{\boldsymbol{\theta}}\widehat{A}],\ \mathcal{J}_{\mathrm{C}}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}_{\mathrm{old}}}[\ell_{a,b}(W_{\boldsymbol{\theta}}, \widehat{A})], \tag{16}$$

and assume that the fixed update advantage obeys $|\widehat{A}| \leq A_{\max}$ almost surely. This bound is operational for standard group normalization: with a population-standard-deviation convention one may take $A_{\max} = \sqrt{G}$, and with an unbiased sample-standard-deviation convention $A_{\max} = \sqrt{G-1}$; adding $\epsilon_A > 0$ only decreases the magnitude. Implementations using a different normalization must state the corresponding constant. Let $\widehat{A}_+ = \max\{\widehat{A}, 0\},\ \widehat{A}_- = \max\{-\widehat{A}, 0\},\ p_{\boldsymbol{\theta}}^+ = \mathbb{P}(W_{\boldsymbol{\theta}} > b),\ p_{\boldsymbol{\theta}}^- = \mathbb{P}(W_{\boldsymbol{\theta}} < a),\ p_{\boldsymbol{\theta}} = p_{\boldsymbol{\theta}}^+ + p_{\boldsymbol{\theta}}^-$. The next proposition bounds the decrease in the surrogate objective caused by clipping.

Proposition 3.3. *For every $\boldsymbol{\theta}$ satisfying the preceding conditions,*

$$0 \leq \mathcal{J}_{\mathrm{G}}(\boldsymbol{\theta}) - \mathcal{J}_{\mathrm{C}}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}_{\mathrm{old}}}\left[\widehat{A}_+(W_{\boldsymbol{\theta}} - b)_+ + \widehat{A}_-(a - W_{\boldsymbol{\theta}})_+\right] \tag{17}$$

$$\leq A_{\max}\left[\sqrt{\left(1 + \frac{2}{L_{\min}}\right)p_{\boldsymbol{\theta}}^+} + a p_{\boldsymbol{\theta}}^-\right] \tag{18}$$

$$\leq A_{\max}\left[\sqrt{\left(1 + \frac{2}{L_{\min}}\right)p_{\boldsymbol{\theta}}} + a p_{\boldsymbol{\theta}}\right]. \tag{19}$$

Objective fidelity does not by itself show that the clipped update points in a useful direction. Define the normalized sequence score vector $\psi_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{L}\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\boldsymbol{y} \mid \boldsymbol{x})$ and assume $\|\psi_{\boldsymbol{\theta}}\| \leq M_\psi$ almost surely, together with the regularity needed to differentiate the unclipped expectation. At $W_{\boldsymbol{\theta}} = a$ or $W_{\boldsymbol{\theta}} = b$, the per-sample clipped objective can be nondifferentiable. We define the CSPO update rule to select the derivative of the unclipped branch at equality, and let g_{C} denote the expectation of this measurable selected direction. Away from the endpoints it equals the classical clipped-objective gradient. An implementation invoking Proposition 3.4 must use the same endpoint derivative convention; otherwise an additional boundary term is needed. Under this convention, the actual saturation event is

$$\mathcal{C}_{\boldsymbol{\theta}} = \{\widehat{A} > 0, W_{\boldsymbol{\theta}} > b\} \cup \{\widehat{A} < 0, W_{\boldsymbol{\theta}} < a\}. \tag{20}$$

Thus $\mathcal{C}_{\boldsymbol{\theta}} \subseteq \{W_{\boldsymbol{\theta}} < a \text{ or } W_{\boldsymbol{\theta}} > b\}$; positive-advantage samples saturate only in the upper tail, and negative-advantage samples only in the lower tail.

Proposition 3.4. *Under the preceding conditions,*

$$g_{\mathrm{G}} := \nabla\mathcal{J}_{\mathrm{G}}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}_{\mathrm{old}}}[W_{\boldsymbol{\theta}}\widehat{A}\psi_{\boldsymbol{\theta}}], \qquad g_{\mathrm{C}} := \mathbb{E}_{\boldsymbol{\theta}_{\mathrm{old}}}[W_{\boldsymbol{\theta}}\widehat{A}\psi_{\boldsymbol{\theta}}\mathbf{1}_{\mathcal{C}_{\boldsymbol{\theta}}^c}], \tag{21}$$

and

$$\|g_{\mathrm{C}} - g_{\mathrm{G}}\| \leq A_{\max} M_\psi \sqrt{\left(1 + \frac{2}{L_{\min}}\right)p_{\boldsymbol{\theta}}}. \tag{22}$$

Moreover, the per-sample clipped gradient $X_{\mathrm{C}} = W_{\boldsymbol{\theta}}\widehat{A}\psi_{\boldsymbol{\theta}}\mathbf{1}_{\mathcal{C}_{\boldsymbol{\theta}}^c}$ satisfies

$$\mathbb{E}_{\boldsymbol{\theta}_{\mathrm{old}}}[\|X_{\mathrm{C}}\|^2] \leq A_{\max}^2 M_\psi^2 \left(1 + \frac{2}{L_{\min}}\right). \tag{23}$$

A sharper diagnostic, which also requires gradient information, is the expected gradient magnitude removed by clipping

$$\tau_{\boldsymbol{\theta}} = \mathbb{E}\left[W_{\boldsymbol{\theta}}|\widehat{A}|\,\|\psi_{\boldsymbol{\theta}}\|\mathbf{1}_{\mathcal{C}_{\boldsymbol{\theta}}}\right], \qquad \|g_{\mathrm{G}} - g_{\mathrm{C}}\| \leq \tau_{\boldsymbol{\theta}}. \tag{24}$$

This inequality follows directly from equation 21. Calibrating $\tau_{\boldsymbol{\theta}}$ would require advantages and gradient norms in addition to score quantiles; the current rule does not control it at level α. The bound on the original normalized score gradient is an assumption that need not hold for a neural network. Clipping individual gradients changes the gradient field and requires a separate bias analysis; clipping the batch gradient does not imply this assumption. Proposition 3.4 therefore requires the stated M_ψ condition or a replacement based on finite moments.

With g_{G} and g_{C} defined in Proposition 3.4, let $B_{\boldsymbol{\theta}} = A_{\max} M_\psi \sqrt{\left(1 + \frac{2}{L_{\min}}\right) p_{\boldsymbol{\theta}}}$. Proposition 3.4 implies $\langle g_{\mathrm{G}}, g_{\mathrm{C}} \rangle \geq \|g_{\mathrm{G}}\|(\|g_{\mathrm{G}}\| - B_{\boldsymbol{\theta}})$. Hence the clipped gradient is an ascent direction for the unclipped power surrogate whenever $B_{\boldsymbol{\theta}} < \|g_{\mathrm{G}}\|$. For a step with $m_t, n_t > 0$ and an update (t, k), the risk entering these fidelity bounds is $p_{t,k}^{\mathrm{band}}$, not its step average. Define

$$D_{t,k} = \|\overline{F}_t - F_{t,k}\|_\infty, \qquad c_{t,k} = \|F_{t,k} - F_t\|_\infty, \qquad D_{t,k} \leq D_t + c_{t,k}. \tag{25}$$

The same two-endpoint argument as in Theorem B.2 gives, for every internal update,

$$p_{t,k}^{\mathrm{band}} \leq \overline{p}_{t,k} := \min\{1, \alpha + 2\varepsilon_t + 2D_{t,k}\} \leq \min\{1, \alpha + 2\varepsilon_t + 2D_t + 2c_{t,k}\}. \tag{26}$$

Indeed, $\|\widehat{F}_{\mathcal{B}_t} - F_{t,k}\|_\infty \leq \varepsilon_t + D_{t,k}$, so each strict tail contributes at most its empirical tail budget plus this discrepancy. Thus $\overline{p}_{t,k}$ may replace $p_{\boldsymbol{\theta}}$ in the objective and gradient bounds at $\boldsymbol{\theta} = \boldsymbol{\theta}_{t,k}$. Control of an entire internal path uses the corresponding bounds for all its iterates.

Corollary B.9 gives the smoothness implication for a subsequent update at a parameter with a valid tail bound. Applying it along a path requires a bound at every iterate; Proposition B.6 states the additional condition for accumulating the gradient error. The following remark specifies the scope of these results.

Remark 3.5. The optimization results compare CSPO with the unclipped GSPO power surrogate. In general, $W_{\boldsymbol{\theta}} = R_{\boldsymbol{\theta}}^{1/L}$ is not a density ratio and

$$\mathbb{E}_{\boldsymbol{\theta}_{\mathrm{old}}}[W_{\boldsymbol{\theta}} f(\boldsymbol{x}, \boldsymbol{y})] \neq \mathbb{E}_{\boldsymbol{\theta}}[f(\boldsymbol{x}, \boldsymbol{y})].$$

Consequently, these results do not imply a KL trust-region constraint or monotonic improvement of the environment reward. Their content is that calibrated tail mass controls how far clipping moves the objective and gradient away from the corresponding GSPO surrogate.

3.2 Pooled bounds and quantile tracking

On the predeclared grid, only the following realized CDF discrepancy is needed:

$$r_t^{\mathcal{G}} = \max_{g \in \mathcal{G}_M} \max\left\{ |\widehat{F}_{\mathcal{B}_t}(g) - \overline{F}_t(g)|, |\widehat{F}_{\mathcal{B}_t}(g^-) - \overline{F}_t(g^-)| \right\}. \tag{27}$$

At the finite grid, define

$$D_{t,\mathcal{G}} = \max_{g \in \mathcal{G}_M} \max\left\{ |\overline{F}_t(g) - F_t(g)|, |\overline{F}_t(g^-) - F_t(g^-)| \right\}. \tag{28}$$

Here $D_{t,\mathcal{G}} \leq D_t \leq \Delta_K$, and $r_t^{\mathcal{G}} \leq \varepsilon_t$ by the weighted triangle inequality.

Theorem 3.6. *For a step t with $m_t, n_t > 0$, let $r_t^{\mathcal{G}}$ and $D_{t,\mathcal{G}}$ be defined above. Use the outward-rounded and anchored endpoints in equation 6–equation 7. Then, on every realized training path,*

$$p_t^{\mathrm{band}} \leq \min\left\{1, \alpha + 2r_t^{\mathcal{G}} + 2D_{t,\mathcal{G}}\right\}. \tag{29}$$

This bound uses the actual pooled score discrepancy and allows dependent, repeatedly scored responses. Corollary B.1 separates its window dependence from linear local drift. Theorem B.10 in Appendix B gives finite-sample rank control for exchangeable scores evaluated at a fixed policy pair. This assumption does not follow for the rolling buffer. Corollary B.11 relates the empirical tail and drift diagnostics to the population discrepancy, and Corollary B.12 records the validity of equation 29 across steps. Proposition B.5 further bounds how closely the endpoints follow population quantiles under the local inverse-CDF condition in Assumption B.4.

4 EXPERIMENTS

We evaluate CSPO on two settings: multi-step embodied decision making using ALFWorld (Shridhar et al., 2021), and difficult mathematical reasoning using DAPO-Math-17k (Yu et al., 2025). ALFWorld is a text-based embodied benchmark in which the agent completes household goals through multi-step interaction. ALFWorld consists of six goal-conditioned task types: Pick & Place (Pick), Examine in Light (Look), Clean & Place (Clean), Heat & Place (Heat), Cool & Place (Cool), and Pick Two & Place (Pick2). For mathematical reasoning, we train on DAPO-Math-17k and evaluate on AIME (MAA, 2024; 2025), AMC (MAA, 2023), MATH500 (Lightman et al., 2024), Minerva Math (Lewkowycz et al., 2022), and OlympiadBench (He et al., 2024). Both settings provide verifiable rewards, obtained directly from the environment for ALFWorld and from answer checking for mathematical reasoning.

We use Qwen3-1.7B for ALFWorld and Qwen3-1.7B-Base for mathematical reasoning (Yang et al., 2025). We compare CSPO with several policy optimization baselines: GRPO (Shao et al., 2024), DAPO (Yu et al., 2025), and GSPO (Zheng et al., 2025). We generate $G = 8$ responses per prompt for mathematical reasoning and $G = 8$ trajectories per task for ALFWorld. Detailed training settings are provided in Appendix C.2.

4.1 MAIN RESULTS

Table 1: Performance on ALFWorld.

Method	ALFWorld						
	Pick	Look	Clean	Heat	Cool	Pick2	All
Qwen3-1.7B	0.00	40.00	3.23	0.00	0.00	0.00	3.91
GSPO	58.06	66.67	45.00	61.54	22.22	20.83	31.25
GRPO	67.86	66.67	56.52	61.54	55.17	**45.45**	52.34
DAPO	76.92	**70.00**	**72.00**	**75.00**	56.52	33.33	53.12
CSPO	**93.33**	**70.00**	68.00	**75.00**	**60.00**	44.00	**64.06**

Table 2: Pass@1 performance on difficult mathematical reasoning benchmarks.

Method	AIME24	AIME25	AMC	MATH500	Minerva	Oly.	Average
Qwen3-1.7B-Base	5.42	3.54	25.38	54.43	13.72	22.47	20.83
GRPO	11.40	5.65	39.43	70.49	21.90	33.27	30.36
DAPO	11.05	5.72	41.56	73.04	22.82	34.32	31.42
GSPO	15.08	8.33	42.86	73.36	22.68	37.83	33.36
CSPO	**16.87**	**8.57**	**43.36**	**74.22**	**23.19**	**38.44**	**34.11**

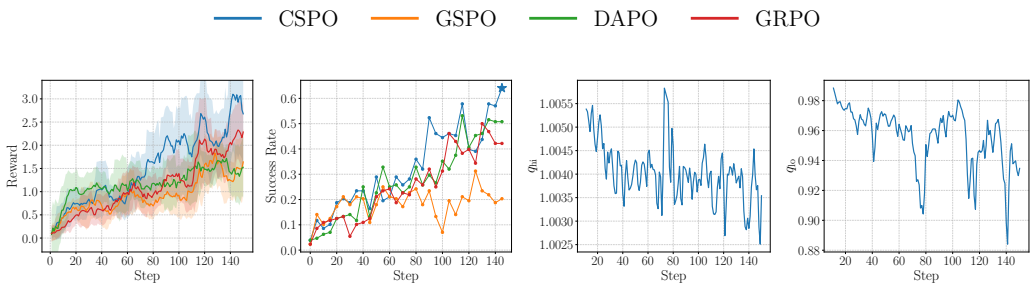
Table 1 reports performance on ALFWorld. CSPO achieves the best overall score of 64.06, improving over DAPO by 10.94 absolute points. CSPO achieves the best or tied-best performance on four of the six task types, including Pick, Look, Heat, and Cool. These results suggest that CSPO is effective for multi-step interactive tasks.

Table 2 reports Pass@1 performance on difficult mathematical reasoning benchmarks. CSPO achieves the best average score of 34.11, improving over GSPO by 0.75 absolute points. Compared with the Qwen3-1.7B-Base model, CSPO improves the average score by 13.28 absolute points. The gains are consistent across AIME, AMC, MATH500, Minerva, and Olympiad-level evaluations, suggesting that adaptive conformal sequence-level clipping is beneficial for difficult reasoning tasks.

4.2 TRAINING DYNAMICS

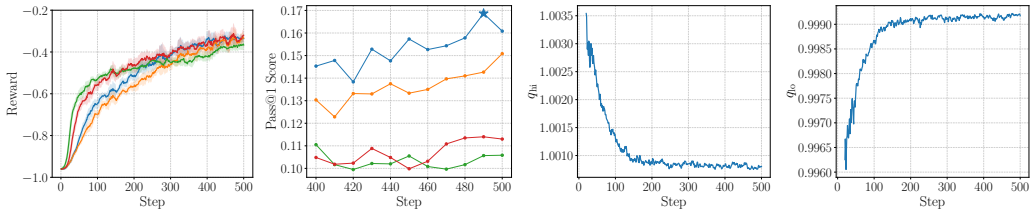
Figure 1 shows training rewards, validation success rates, and the calibrated endpoints on ALFWorld. The endpoints fluctuate throughout training as the distribution of retained scores changes. The upper and lower endpoints follow different trajectories, allowing the interval to adapt separately to the two tails.

Figure 2 shows the corresponding curves for DAPO-Math-17k, including Pass@1 on AIME24. During early training, the upper endpoint decreases toward one and the lower endpoint increases toward one; both then vary within a narrower range. The band thus contracts as the retained power scores become more concentrated. These curves describe the behavior of the adaptive band; assessing its population tail probability also requires the CDF discrepancies in Section 3.



(a) Reward curves. (b) Success rate. (c) Dynamics of q_{hi}. (d) Dynamics of q_{lo}.

Figure 1: Learning curves and evaluation results on ALFWorld.



(a) Reward curves. (b) Pass@1 on AIME24. (c) Dynamics of q_{hi}. (d) Dynamics of q_{lo}.

Figure 2: Learning curves and evaluation results on DAPO-Math-17k.

Sensitivity and computational cost. The main experiments use $\alpha = 0.1$ and $K = 2$, selected using preliminary ablations on MATH and GSM8K (Appendix C.3). On DAPO-Math-17k, increasing α from 0.1 to 0.2 and 0.4 reduces average Pass@1 from 34.11 to 33.50 and 32.37, respectively (Table 5). For a fixed buffer and proportional allocation between the tails, a smaller α widens the quantile interval and leaves more scores inside the band. The reported results favor $\alpha = 0.1$ among the tested values. Figure 6 in Appendix C.4 shows comparable training-step and actor-update times for CSPO and the baselines.

5 Conclusion

CSPO adapts sequence-level clipping bounds using quantiles of recent training scores. Its tail bounds account for data reuse and policy drift, and its per-update bounds control the objective and gradient error under the stated moment conditions. In our experiments, CSPO improves over GSPO on mathematical reasoning benchmarks and achieves the highest overall success rate on ALFWorld. The changing endpoints provide a direct view of how the clipping band adapts during training.

Limitations. The population tail bounds depend on discrepancies between the CDFs of retained scores and their population mixtures, as well as local CDF drift. Reusing a rollout batch does not yield a sampling rate for these discrepancies from the number of retained entries alone. The gradient bound also assumes a bound on the normalized score gradient that need not hold for a neural network. The results quantify error relative to the GSPO power surrogate; they do not establish a KL constraint or reward improvement.

Reproducibility statement

Algorithm 1 specifies score retention and buffer updates; Section 3 states the assumptions; and Appendix B provides complete proofs. Appendix C.2 records the training settings, and Appendix C.3 reports the choices of α and K and their ablations.

AI use statement

Generative AI tools were used to aid or polish writing.

References

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, 2017.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Isaac Gibbs and Emmanuel J. Candès. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems*, 2021.

Isaac Gibbs and Emmanuel J. Candès. Conformal inference for online prediction with arbitrary distribution shifts. *Journal of Machine Learning Research*, 25(162):1–36, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li,

Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 9 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL https://doi.org/10.1038/s41586-025-09422-z.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. In *Annual Meeting of the Association for Computational Linguistics*, 2024.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving quantitative reasoning problems with language models. *arXiv preprint arXiv:2206.14858*, 2022.

Yuan Li, Bo Wang, Yufei Gao, Yuqian Yao, Xinyuan Wang, Zhangyue Yin, and Xipeng Qiu. BandPO: Bridging trust regions and ratio clipping via probability-aware bounds for LLM reinforcement learning. *arXiv preprint arXiv:2603.04918*, 2026. doi: 10.48550/arXiv.2603.04918.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024.

MAA. American mathematics contest 12 (amc 12). https://artofproblemsolving.com/wiki/index.php/AMC_12_Problems_and_Solutions, November 2023.

MAA. American invitational mathematics examination (aime). https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions, February 2024.

MAA. American invitational mathematics examination (aime). https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions, February 2025.

Hanyi Mao, Quanjia Xiao, Lei Pang, and Haixiao Liu. Clip your sequences fairly: Enforcing length fairness for sequence-level rl. *arXiv preprint arXiv:2509.09177*, 2025.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, 2022.

Drew Prinster, Clara Fannjiang, Ji Won Park, Kyunghyun Cho, Anqi Liu, Suchi Saria, and Samuel Stanton. Conformal policy control. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, 2023.

Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, 2019.

John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. Trust region policy optimization. In *International Conference on Machine Learning*, pp. 1889–1897, 2015.

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

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421, 2008.

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

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Cote, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. {ALFW}orld: Aligning text and embodied environments for interactive learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=0IOX0YcCdTn.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. Learning to summarize from human feedback. In *Advances in Neural Information Processing Systems*, 2020.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Peidong Wang, Demi Wang, Xufang Luo, Jiahang Xu, Xiaocui Yang, Shi Feng, Yuqing Yang, and Dongsheng Li. What are key factors for updates in rl for llm reasoning? *arXiv preprint arXiv:2606.22570*, 2026.

Xumeng Wen, Zihan Liu, Shun Zheng, Zhijian Xu, Shengyu Ye, Zhirong Wu, Xiao Liang, Yang Wang, Junjie Li, Ziming Miao, Jiang Bian, and Mao Yang. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base llms. *arXiv preprint arXiv:2506.14245*, 2025.

Zhiheng Xi, Xin Guo, Yang Nan, Enyu Zhou, Junrui Shen, Wenxiang Chen, Jiaqi Liu, Jixuan Huang, Zhihao Zhang, Honglin Guo, Xun Deng, Zhikai Lei, Miao Zheng, Guoteng Wang, Shuo Zhang, Peng Sun, Rui Zheng, Hang Yan, Tao Gui, Qi Zhang, and Xuanjing Huang. BAPO: Stabilizing off-policy reinforcement learning for LLMs via balanced policy optimization with adaptive clipping. *arXiv preprint arXiv:2510.18927*, 2025. doi: 10.48550/arXiv.2510.18927.

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

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan 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, Mu Qiao, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale, 2025. URL https://arxiv.org/abs/2503.14476.

Yuheng Zhang, Chenlu Ye, Shuowei Jin, Changlong Yu, Wei Xiong, Saurabh Sahu, and Nan Jiang. Rethinking importance sampling in LLM policy optimization: A cumulative token perspective. *arXiv preprint arXiv:2605.07331*, 2026. doi: 10.48550/arXiv.2605.07331.

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. *arXiv preprint arXiv:2507.18071*, 2025.

A Related Work

Related work relevant to CSPO spans three complementary directions: reinforcement learning for language-model post-training, the transition from token-level to sequence-level policy optimization, and conformal prediction and adaptive calibration. We review these strands in turn to position CSPO as a sequence-level policy optimization method that replaces a prespecified clipping band with one calibrated from retained training scores.

Reinforcement learning for language model post-training. RL-based post-training has a long history in aligning models with human or task-specific preferences. Early preference-based RL showed that complex objectives can be learned from pairwise human comparisons (Christiano et al., 2017). In language modeling, RLHF pipelines train reward models from human preference data and optimize policies to maximize the learned reward while constraining deviation from a reference model (Stiennon et al., 2020; Ouyang et al., 2022). Constitutional AI reduces direct human labeling by using principle-guided AI feedback for harmlessness training (Bai et al., 2022). In parallel, direct preference optimization methods bypass explicit reward-model RL by deriving a supervised objective from the KL-regularized RLHF optimum (Rafailov et al., 2023). These methods focus primarily on preference alignment, while our work targets policy optimization for verifiable reasoning rewards.

RLVR has recently become a central approach for improving mathematical and coding reasoning. When rewards are automatically verifiable, the supervision signal can be scaled without dense human annotations. DeepSeekMath introduced GRPO to improve mathematical reasoning while reducing PPO memory overhead (Shao et al., 2024). DeepSeek-R1 further demonstrated that reasoning behaviors such as reflection and verification can emerge through large-scale RL (Guo et al., 2025). DAPO highlights several practical ingredients for large-scale long-CoT RL, including decoupled clipping and dynamic sampling (Yu et al., 2025). Our method is complementary to these systems: CSPO modifies the clipping geometry of the policy objective and can be combined with different reward functions, sampling strategies, and KL regularizers.

From token-level to sequence-level policy optimization. PPO stabilizes policy-gradient updates by clipping the probability ratio at the action level (Schulman et al., 2017). In autoregressive language models, applying this design token-by-token is convenient, but the reward and advantage are often defined for an entire generated response. GRPO keeps this token-level clipping structure while replacing the value model with group-relative sequence rewards (Shao et al., 2024). This design is efficient, but it implicitly assigns a single response-level advantage to many token-level ratios. As a result, the objective may over-penalize locally off-policy tokens or distort the contribution of a high-value response.

Sequence-level methods address this mismatch by moving the clipping multiplier to the response level. GSPO constructs a power multiplier from the sequence likelihood ratio and performs sequence-level clipping and optimization (Zheng et al., 2025). FSPO further studies the length fairness problem of sequence-level clipping and shows that fixed clipping bands can reweight short and long responses differently; it introduces a $\sqrt{L}$-scaled clipping band for the sequence log-ratio (Mao et al., 2025). CSPO extends this approach by calibrating the clipping band from the recent empirical distribution of normalized sequence scores, allowing the endpoints to change during training.

Conformal prediction and adaptive calibration. Conformal prediction provides distribution-free uncertainty quantification by calibrating prediction sets using held-out nonconformity scores (Vovk et al., 2005; Shafer & Vovk, 2008; Angelopoulos & Bates, 2023). Its central guarantee is finite-sample marginal coverage under exchangeability, independent of the internal form of the predictive model. Conformalized quantile regression combines conformal calibration with quantile regression to obtain adaptive prediction intervals under heteroscedasticity (Romano et al., 2019). Adaptive conformal inference extends calibration to online environments where the data distribution may vary over time (Gibbs & Candès, 2021; 2024).

We use quantile calibration to assign a tail budget to fresh sequence-level scores outside the clipping band. The GSPO power score $W = \exp(S/L)$ plays the role of the calibration score: values in the tails indicate unusually large policy shifts. By maintaining a rolling training-score buffer, CSPO adapts the clipping interval to the current training regime while targeting a marginal tail budget with explicit CDF discrepancies and local drift.

B THEORETICAL DETAILS

This appendix states the supplementary corollaries used in the main discussion and verifies all formal results under the same score, sampling unit, thresholds, and conditioning used by the main algorithm.

B.1 AUTOREGRESSIVE POLICY AND GRPO OBJECTIVE

For completeness, this subsection records the autoregressive factorization and token-level reference objective that are not part of CSPO's technical novelty. For $\boldsymbol{y}_i = (y_{i,1}, \ldots, y_{i,L_i})$ and $\boldsymbol{h}_{i,\ell} = (\boldsymbol{x}_i, \boldsymbol{y}_{i,<\ell})$,

$$\pi_{\boldsymbol{\theta}}(\boldsymbol{y}_i \mid \boldsymbol{x}_i) = \prod_{\ell=1}^{L_i} \pi_{\boldsymbol{\theta}}(y_{i,\ell} \mid \boldsymbol{h}_{i,\ell}). \tag{30}$$

For G responses, use the group-normalized $\widehat{A}_i$ and token ratio $r_{i,\ell}(\boldsymbol{\theta})$ defined in Section 2. The token-level GRPO reference is

$$J_{\mathrm{tok}}(\boldsymbol{\theta}) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{L_i}\sum_{\ell=1}^{L_i}\min\Big\{r_{i,\ell}(\boldsymbol{\theta})\widehat{A}_i, \mathrm{clip}(r_{i,\ell}(\boldsymbol{\theta}), 1-\epsilon, 1+\epsilon)\widehat{A}_i\Big\}\right]. \tag{31}$$

B.2 DETAILED METHOD POSITIONING

Table 3 separates the object used to construct a clipping rule from the object controlled by its stated analysis.

Table 3: Comparison with representative clipping and conformal-control methods. "Controlled object" denotes the stated construction or guarantee, not an empirical performance claim.

Method	Update element or score	Band/control construction	Primary controlled object
PPO / GRPO	Token ratio	Fixed numerical endpoints	Pessimistic token surrogate
GSPO	Response $W = e^{S/L}$	Fixed response-level endpoints	Sequence-coherent power surrogate
FSPO	Sequence log-ratio	Fixed $\sqrt{L}$-scaled geometry	Length reweighting and gradient direction
CTPO (Zhang et al., 2026)	Prefix/cumulative ratio	Position-dependent scaling	Prefix correction and position scaling
BandPO (Li et al., 2026)	Probability ratio	Divergence-derived probability-aware bounds	Projected trust-region geometry
BAPO (Xi et al., 2025)	Ratio contributions	Advantage/entropy-driven asymmetry	Contribution balance and entropy
Adaptive Clip PO (Wang et al., 2026)	Token-group ratios	Empirical-variance endpoint scaling	Gradient/update heterogeneity
Conformal Policy Control (Prinster et al., 2026)	Policy relative to safe reference	Conformal control-parameter selection	Deployment-time safety loss
CSPO	Response $W = e^{S/L}$	Training-score quantiles, outward grid, and drift	Fresh-score tail mass and GSPO-surrogate fidelity

B.3 BUFFER SIZE AND LOCAL DRIFT

For a window of length K, write $r^{\mathcal{G}}_{t,K}$ for its realized grid discrepancy.

Corollary B.1. *Suppose the conditions of Theorem 3.6 hold and the preceding K steps have equal numbers m of cached score entries. If, for some $\rho \geq 0$,*

$$\|F_{t-r} - F_t\|_\infty \leq \rho r, \qquad r = 1, \ldots, K,$$

then, pathwise,

$$p_t^{\mathrm{band}} \leq \min\Big\{1, \alpha + 2r^{\mathcal{G}}_{t,K} + \rho(K+1)\Big\}. \tag{32}$$

For a finite set $\mathcal{K}_t$ of available window lengths satisfying the same drift envelope, the corresponding oracle comparison is

$$K_{\mathrm{oracle}} \in \arg\min_{K \in \mathcal{K}_t}\left\{2r^{\mathcal{G}}_{t,K} + \rho(K+1)\right\}. \tag{33}$$

The discrepancy $r^{\mathcal{G}}_{t,K}$ need not decrease with K; this oracle comparison is not a data-only window-selection rule.

B.4 SECONDARY TRANSFER, TRACKING, AND PATH RESULTS

This subsection bounds the raw quantile tail probability, extends the result to length groups, and analyzes quantile tracking and error accumulation across updates. The tail bound allows discrete scores and dependence across training steps. Rounding and anchoring can only reduce the probability of falling outside the band.

Theorem B.2. *For a step t with $m_t, n_t > 0$, define*

$$\varepsilon_t = \sum_{\tau \in \mathcal{I}_t} \frac{m_\tau}{n_t} \xi_\tau, \qquad \xi_\tau = \|\widehat{F}_\tau - F_\tau\|_\infty. \tag{34}$$

Then the first two bounds below hold on every realized training path; the last also uses Assumption 3.1:

$$\begin{aligned} p_t^{\text{band}} &:= \mathbb{P}(W_t^\star < a_t \text{ or } W_t^\star > b_t \,|\, \mathcal{H}_t) \\ &\le p_t^q := \mathbb{P}(W_t^\star < q_{\text{lo},t} \text{ or } W_t^\star > q_{\text{hi},t} \,|\, \mathcal{H}_t) \\ &\le \alpha + 2\varepsilon_t + 2D_t \le \alpha + 2\varepsilon_t + 2\Delta_K. \end{aligned} \tag{35}$$

Theorem B.2 is intentionally one-sided: deterministic quantiles can be conservative when an endpoint has large probability mass. Let $\omega_t = \sup_w\{F_t(w) - F_t(w^-)\}$ be the largest current atom. Corollary B.7 shows that the same path also satisfies

$$|p_t^q - \alpha| \le 2\big(\varepsilon_t + D_t\big) + 2\omega_t.$$

The upper bound allows ties under local CDF drift, while closeness to the target from both sides also depends on endpoint mass. The marginal theorem averages over the response-length distribution. To determine when it also describes a length subgroup, let $H_t^\star = h(L_t^\star)$ be a pre-specified length bin with positive conditional probability and define

$$F_{t,h}(w) = \mathbb{P}(W_t^\star \le w \,|\, \mathcal{H}_t, H_t^\star = h), \qquad \Gamma_{t,h} = \sup_w |F_{t,h}(w) - F_t(w)|.$$

The quantity $\Gamma_{t,h}$ measures the CDF discrepancy between length group h and the full mixture; normalization by L alone does not force it to be small.

Proposition B.3. *Let*

$$p_{t,h}^{\text{band}} = \mathbb{P}(W_t^\star < a_t \text{ or } W_t^\star > b_t \,|\, \mathcal{H}_t, H_t^\star = h).$$

Then, almost surely,

$$|p_{t,h}^{\text{band}} - p_t^{\text{band}}| \le 2\Gamma_{t,h}. \tag{36}$$

Combining Proposition B.3 with Theorem B.2 gives the group upper bound

$$p_{t,h}^{\text{band}} \le \alpha + 2\big(\varepsilon_t + D_t\big) + 2\Gamma_{t,h}.$$

Corollary B.8 also gives the atom-corrected two-sided statement when anchoring is inactive.

These results separate marginal calibration from length-stratified calibration. A small CDF discrepancy $\Gamma_{t,h}$ supports transfer to group h. Coverage alone does not quantify how closely the rolling band follows the current quantiles. The next assumption adds the local identifiability needed to convert a CDF error into endpoint error; it is not required by Theorems B.2 or 3.6.

Assumption B.4. For $p \in \{\alpha_{\text{lo}}, 1 - \alpha_{\text{hi}}\}$ and every analyzed round, let $q_{t,p}^\star = F_t^{-1}(p) := \inf\{w : F_t(w) \ge p\}$. There exist $\underline{f} > 0$ and $r_q > 0$ such that F_t is continuous on $[q_{t,p}^\star - r_q, q_{t,p}^\star + r_q]$ and

$$|F_t(q_{t,p}^\star + u) - p| \ge \underline{f}|u|, \qquad |u| \le r_q.$$

Every such neighborhood lies strictly inside the finite grid range. For a predeclared resolution $h_{\mathcal{G}} \ge 0$, every z in such a neighborhood has grid points $g^-(z), g^+(z) \in \mathcal{G}_M$ satisfying

$$g^-(z) \le z \le g^+(z), \qquad z - g^-(z) \le h_{\mathcal{G}}, \qquad g^+(z) - z \le h_{\mathcal{G}}.$$

Along each analyzed path, the CDF discrepancies below obey $r_t^{\mathcal{G}} + D_t < \underline{f} r_q$.

Assumption B.4 excludes flat CDF regions around the target quantiles. Under this condition, the following result bounds the cumulative error in tracking the quantiles of the score mixture.

Proposition B.5. *Suppose Assumptions 3.1 and B.4 hold for rounds $t = K+1, \dots, T$, with equal block size m, a fixed horizon K, and the grid endpoints used by Algorithm 1. Let*

$$v_t = \|F_t - F_{t-1}\|_\infty, \qquad V_T = \sum_{t=2}^{T} v_t.$$

Set $r^{\mathcal{G}}_{T,K} = \max_{K<t\le T} r^{\mathcal{G}}_t$. For either mixture quantile level p, let $\widetilde{q}_{t,p}$ be the corresponding outward-rounded empirical endpoint. On every path satisfying the stated conditions,

$$\sum_{t=K+1}^{T} |\widetilde{q}_{t,p} - q^\star_{t,p}| \le (T-K)h_{\mathcal{G}} + \frac{1}{\underline{f}}\left[(T-K)r^{\mathcal{G}}_{T,K} + \frac{K+1}{2}V_T\right]. \tag{37}$$

If $\mathbb{E}_{F_t}[W] < \infty$ for every analyzed round, then for the pinball risk $\mathcal{L}_{t,p}(q) = \mathbb{E}_{F_t}[\rho_p(W-q)]$, $\rho_p(u) = u(p - \mathbf{1}\{u<0\})$, the cumulative dynamic regret $\sum_t[\mathcal{L}_{t,p}(\widetilde{q}_{t,p}) - \mathcal{L}_{t,p}(q^\star_{t,p})]$ is bounded by the same right-hand side.

Let $(a^\star_t, b^\star_t)$ be the oracle quantile endpoints anchored at 1. Let $\mathcal{J}_{\mathrm{G},t} = \sum_{k\in\mathcal{C}_t} \nu_{t,k}\mathcal{J}_{\mathrm{G}}(\boldsymbol{\theta}_{t,k})$ and $\mathcal{J}^{a,b}_{\mathrm{C},t} = \sum_{k\in\mathcal{C}_t} \nu_{t,k}\mathcal{J}^{a,b}_{\mathrm{C}}(\boldsymbol{\theta}_{t,k})$ be the corresponding collection-weighted population surrogates, all with denominator $\pi_{\boldsymbol{\theta}_{t,0}}$. Define $\mathfrak{D}_t(a,b) = \mathcal{J}_{\mathrm{G},t} - \mathcal{J}^{a,b}_{\mathrm{C},t}$, keeping these weights fixed when comparing bands. If, in addition, the finite-endpoint and moment conditions of Proposition 3.3 hold at every collected update and $|\widehat{A}| \le A_{\max}$, then

$$\sum_{t=K+1}^{T} |\mathfrak{D}_t(a_t,b_t) - \mathfrak{D}_t(a^\star_t, b^\star_t)| \le 2A_{\max}\left\{(T-K)h_{\mathcal{G}} + \frac{(T-K)r^{\mathcal{G}}_{T,K} + (K+1)V_T/2}{\underline{f}}\right\}. \tag{38}$$

The dynamic regret in Proposition B.5 establishes that CSPO follows its chosen quantile target under drift; it does not claim that this target maximizes environment reward. In particular, an unbounded band trivially has zero clipping distortion but abandons the intended regularization budget. The empirical evaluation must therefore report both calibration fidelity and task performance.

The preceding results operate across outer rounds. To state what tail control would imply within one rollout step, we finally compare two exact-gradient paths for one fixed old policy, advantage law, and surrogate. In this path comparison, the fixed outer-step index is suppressed. This result also identifies the missing condition: a bound at one parameter value does not certify every inner iterate.

Proposition B.6. *Let*

$$\boldsymbol{\theta}^{\mathrm{G}}_{k+1} = \boldsymbol{\theta}^{\mathrm{G}}_k + \gamma_k g_{\mathrm{G}}(\boldsymbol{\theta}^{\mathrm{G}}_k), \qquad \boldsymbol{\theta}^{\mathrm{C}}_{k+1} = \boldsymbol{\theta}^{\mathrm{C}}_k + \gamma_k g_{\mathrm{C}}(\boldsymbol{\theta}^{\mathrm{C}}_k), \qquad \boldsymbol{\theta}^{\mathrm{G}}_0 = \boldsymbol{\theta}^{\mathrm{C}}_0.$$

Suppose $\gamma_k \ge 0$, $B_k \ge 0$, g_{G} is L_g-Lipschitz on a set containing both paths, and $\|g_{\mathrm{C}}(\boldsymbol{\theta}^{\mathrm{C}}_k) - g_{\mathrm{G}}(\boldsymbol{\theta}^{\mathrm{C}}_k)\| \le B_k$. Then, for every $E \ge 1$,

$$\|\boldsymbol{\theta}^{\mathrm{C}}_E - \boldsymbol{\theta}^{\mathrm{G}}_E\| \le \sum_{j=0}^{E-1} \gamma_j B_j \prod_{\ell=j+1}^{E-1} (1 + \gamma_\ell L_g). \tag{39}$$

If Proposition 3.4 is valid at every CSPO iterate and p_k is the fresh-score risk at $\boldsymbol{\theta}^{\mathrm{C}}_k$ relative to the same fixed old policy, band, and advantage law, one may take $B_k = A_{\max}M_\psi\sqrt{(1+2/L_{\min})p_k}$. This substitution requires the absolute-continuity, length, advantage, and score-gradient conditions at every iterate, together with a simultaneous bound on the risks $(p_k)_k$ along the path. The statement concerns deterministic exact-gradient paths, not an SGD trajectory.

B.5 Supplementary calibration and optimization corollaries

The main theorem gives the tie-robust upper guarantee used by the operational algorithm. To characterize when the raw quantile interval is also close to the nominal tail mass from below, the following corollary makes endpoint atoms explicit.

Corollary B.7. *Under the conditions of Theorem B.2, define*

$$\omega_t = \sup_{w\in\mathbb{R}}\{F_t(w) - F_t(w^-)\}, \qquad F_t(w^-) = \lim_{u\uparrow w} F_t(u). \tag{40}$$

On the same realized path,

$$\alpha - 2\big(\varepsilon_t + D_t\big) - 2\omega_t \le p_t^q \le \alpha + 2\big(\varepsilon_t + D_t\big). \tag{41}$$

Consequently,

$$|p_t^q - \alpha| \le 2\big(\varepsilon_t + D_t\big) + 2\omega_t. \tag{42}$$

The length-transfer proposition isolates the discrepancy between a group CDF and the marginal CDF. Combining it with the calibration theorem yields the following operational group guarantee.

Corollary B.8. *Under the conditions of Theorem B.2, every length group h with positive conditional probability satisfies the following pathwise bound:*

$$p_{t,h}^{\mathrm{band}} \le \alpha + 2\big(\varepsilon_t + D_t\big) + 2\Gamma_{t,h}. \tag{43}$$

If rounding and anchoring leave the raw interval unchanged, so that $(a_t, b_t) = (q_{\mathrm{lo},t}, q_{\mathrm{hi},t})$, then

$$|p_{t,h}^{\mathrm{band}} - \alpha| \le 2\big(\varepsilon_t + D_t\big) + 2\omega_t + 2\Gamma_{t,h}. \tag{44}$$

The gradient-fidelity proposition gives a positive inner product when the unclipped gradient dominates its calibrated tail error. The following auxiliary statement applies when its gradient estimator is conditionally unbiased and independent across prompt groups; these properties are not inferred for a reused-batch update at $k > 0$.

Corollary B.9. *Let $\mathcal{J}$ be a sigma-field fixing the parameter pair and band before the gradient-estimator randomness. Assume $L_{\mathrm{s}} > 0$ and that $\mathcal{J}_{\mathrm{G}}$ is L_{s}-smooth on a convex set containing $\boldsymbol{\theta}$ and $\boldsymbol{\theta} + \gamma\widehat{g}_{\mathrm{C}}$ almost surely for every step size considered. Let*

$$C_{\mathrm{grad}} = A_{\max}^2 M_\psi^2 \left(1 + \frac{2}{L_{\min}}\right).$$

For prompt group j, let $X_{j,i}$ denote the per-response selected direction X_{C} from Proposition 3.4 and define

$$Z_j = \frac{1}{G}\sum_{i=1}^{G} X_{j,i}.$$

Let $Z_1, \dots, Z_N$ be conditionally independent across prompt groups given $\mathcal{J}$, with $\mathbb{E}[Z_j \mid \mathcal{J}] = g_{\mathrm{C}}$, and set $\widehat{g}_{\mathrm{C}} = N^{-1}\sum_{j=1}^{N} Z_j$. Then

$$\mathbb{E}[\|\widehat{g}_{\mathrm{C}} - g_{\mathrm{C}}\|^2 \mid \mathcal{J}] \le \frac{C_{\mathrm{grad}}}{N}.$$

For $\boldsymbol{\theta}^+ = \boldsymbol{\theta} + \gamma\widehat{g}_{\mathrm{C}}$,

$$\mathbb{E}\big[\mathcal{J}_{\mathrm{G}}(\boldsymbol{\theta}^+) - \mathcal{J}_{\mathrm{G}}(\boldsymbol{\theta}) \,\big|\, \mathcal{J}\big] \ge \gamma\|g_{\mathrm{G}}\|(\|g_{\mathrm{G}}\| - B_{\boldsymbol{\theta}}) - \frac{L_{\mathrm{s}}\gamma^2}{2}\left[(\|g_{\mathrm{G}}\| + B_{\boldsymbol{\theta}})^2 + \frac{C_{\mathrm{grad}}}{N}\right]. \tag{45}$$

If $B_{\boldsymbol{\theta}} < \|g_{\mathrm{G}}\|$, the lower bound is positive for every

$$0 < \gamma < \frac{2\|g_{\mathrm{G}}\|(\|g_{\mathrm{G}}\| - B_{\boldsymbol{\theta}})}{L_{\mathrm{s}}[(\|g_{\mathrm{G}}\| + B_{\boldsymbol{\theta}})^2 + C_{\mathrm{grad}}/N]}. \tag{46}$$

The rolling theorem is intentionally nonexchangeable. To identify the stronger conformal benchmark precisely, consider instead calibration and future scores evaluated by one policy pair that was fixed before either was sampled. The rank construction below is needed only for this stationary reference case and is not used to justify pooling across policy pairs.

Theorem B.10. *Condition on a sigma-field $\mathcal{G}$ with respect to which a policy pair and its score map are fixed. Suppose positive scores $W_1, \dots, W_n, W^\star$ are exchangeable conditional on $\mathcal{G}$. Let $W_{(1)} \le \cdots \le W_{(n)}$, set $W_{(0)} = 0$ and $W_{(n+1)} = +\infty$, and define*

$$k_{\mathrm{lo}} = \lfloor (n+1)\alpha_{\mathrm{lo}} \rfloor, \qquad k_{\mathrm{hi}} = \lceil (n+1)(1-\alpha_{\mathrm{hi}}) \rceil.$$

With

$$q_{\mathrm{lo}} = W_{(k_{\mathrm{lo}})}, \qquad q_{\mathrm{hi}} = W_{(k_{\mathrm{hi}})}, \qquad a = \min\{q_{\mathrm{lo}}, 1\}, \qquad b = \max\{q_{\mathrm{hi}}, 1\},$$

one has, for arbitrary discrete or continuous score laws,

$$\mathbb{P}(W^\star < a \text{ or } W^\star > b \mid \mathcal{G}) \le \frac{k_{\mathrm{lo}}}{n+1} + \frac{n+1-k_{\mathrm{hi}}}{n+1} \le \alpha. \tag{47}$$

Strict tail events make ties conservative. Equality at the nominal level requires randomized tie and boundary-rank handling and is not claimed here.

The current completed score block permits an empirical comparison with the retained blocks.

Corollary B.11. *For $m_t, n_t > 0$, define the empirical tail probability for the step mixture and the drift diagnostic*

$$\widehat{p}_t^{\,\mathrm{band}} = \widehat{F}_t(a_t^-) + 1 - \widehat{F}_t(b_t), \qquad \widehat{D}_t = \sum_{\tau \in \mathcal{I}_t} \frac{m_\tau}{n_t} \|\widehat{F}_\tau - \widehat{F}_t\|_\infty. \tag{48}$$

These use the actual cached entries, including repeated entries. For the quantile band and its outward-rounded, anchored version, pathwise,

$$\widehat{p}_t^{\,\mathrm{band}} \le \min\{1, \alpha + 2\widehat{D}_t\}, \qquad p_t^{\mathrm{band}} \le \min\{1, \alpha + 2\widehat{D}_t + 2\xi_t\}. \tag{49}$$

Also $D_t \le \widehat{D}_t + \varepsilon_t + \xi_t$. If $\widehat{\omega}_t = \sup_w\{\widehat{F}_t(w) - \widehat{F}_t(w^-)\}$, then

$$\omega_t \le \min\{1, \widehat{\omega}_t + 2\xi_t\}. \tag{50}$$

The empirical bound is directly computable. The population and atom bounds additionally contain the realized discrepancy ξ_t; they are not data-only confidence bounds.

Corollary B.12. *Along any training path, simultaneously at every step with $m_t, n_t > 0$,*

$$p_t^{\mathrm{band}} \le \min\{1, \alpha + 2r_t^{\mathcal{G}} + 2D_{t,\mathcal{G}}\}. \tag{51}$$

The same inequality therefore holds at a finite step selected from the observed training history, retaining that step's realized discrepancy and drift.

B.6 Proof of Theorem B.2

Proof. Fix the realized training path. By equation 10, the empirical and population step CDFs use the same collection weights, so

$$\widehat{F}_\tau - F_\tau = \sum_{k \in \mathcal{C}_\tau} \nu_{\tau,k}(\widehat{F}_{\tau,k} - F_{\tau,k}), \qquad \xi_\tau \le \sum_{k \in \mathcal{C}_\tau} \nu_{\tau,k}\xi_{\tau,k}.$$

This identity holds for repeated or dependent entries as well as for distinct entries.

Let

$$\lambda_{\tau,t} = \frac{m_\tau}{n_t}, \qquad \overline{F}_t(w) = \sum_{\tau \in \mathcal{I}_t} \lambda_{\tau,t} F_\tau(w).$$

Since

$$\widehat{F}_{\mathcal{B}_t}(w) = \sum_{\tau \in \mathcal{I}_t} \lambda_{\tau,t} \widehat{F}_\tau(w),$$

pathwise,

$$\sup_w |\widehat{F}_{\mathcal{B}_t}(w) - \overline{F}_t(w)| \le \sum_{\tau \in \mathcal{I}_t} \lambda_{\tau,t}\xi_\tau = \varepsilon_t.$$

When Assumption 3.1 holds, it additionally gives

$$\sup_w |\overline{F}_t(w) - F_t(w)| = D_t \le \sum_{\tau \in \mathcal{I}_t} \lambda_{\tau,t} \sup_w |F_\tau(w) - F_t(w)| \le \Delta_K.$$

Thus, on the same path,

$$\sup_w |\widehat{F}_{\mathcal{B}_t}(w) - F_t(w)| \le \eta_t, \qquad \eta_t := \varepsilon_t + D_t. \tag{52}$$

By the definition of the empirical inverse CDF,

$$\widehat{F}_{\mathcal{B}_t}(q_{\mathrm{lo},t}^-) \le \alpha_{\mathrm{lo}}, \qquad \widehat{F}_{\mathcal{B}_t}(q_{\mathrm{hi},t}) \ge 1 - \alpha_{\mathrm{hi}}.$$

Taking left limits in equation 52 is valid for CDFs, so

$$F_t(q_{\mathrm{lo},t}^-) \le \alpha_{\mathrm{lo}} + \eta_t, \qquad 1 - F_t(q_{\mathrm{hi},t}) \le \alpha_{\mathrm{hi}} + \eta_t.$$

The empirical quantiles are ordered because $\alpha_{\mathrm{lo}} < 1 - \alpha_{\mathrm{hi}}$. Therefore the strict tail events are disjoint and

$$p_t^q = F_t(q_{\mathrm{lo},t}^-) + 1 - F_t(q_{\mathrm{hi},t}) \le \alpha + 2\eta_t.$$

Finally, $a_t \le q_{\mathrm{lo},t}$ and $b_t \ge q_{\mathrm{hi},t}$, so

$$\{W_t^\star < a_t \text{ or } W_t^\star > b_t\} \subseteq \{W_t^\star < q_{\mathrm{lo},t} \text{ or } W_t^\star > q_{\mathrm{hi},t}\}.$$

This proves both inequalities in equation 35. □

B.7 Proof of Corollary B.7

Proof. Retain $\eta_t = \varepsilon_t + D_t$ from the preceding pathwise proof. The inverse empirical quantiles also satisfy

$$\widehat{F}_{\mathcal{B}_t}(q_{\mathrm{lo},t}) \ge \alpha_{\mathrm{lo}}, \qquad \widehat{F}_{\mathcal{B}_t}(q_{\mathrm{hi},t}^-) \le 1 - \alpha_{\mathrm{hi}}.$$

Using equation 52,

$$F_t(q_{\mathrm{lo},t}) \ge \alpha_{\mathrm{lo}} - \eta_t, \qquad F_t(q_{\mathrm{hi},t}^-) \le 1 - \alpha_{\mathrm{hi}} + \eta_t.$$

The mass at either endpoint is at most ω_t. Hence

$$F_t(q_{\mathrm{lo},t}^-) \ge \alpha_{\mathrm{lo}} - \eta_t - \omega_t$$

and

$$1 - F_t(q_{\mathrm{hi},t}) \ge \alpha_{\mathrm{hi}} - \eta_t - \omega_t.$$

Adding these bounds gives $p_t^q \ge \alpha - 2\eta_t - 2\omega_t$. The upper bound follows from Theorem B.2, and the absolute-deviation statement follows by combining the two sides. □

B.8 Proof of Proposition B.3

Proof. Conditional on $\mathcal{H}_t$,

$$p_t^{\mathrm{band}} = F_t(a_t^-) + 1 - F_t(b_t)$$

and

$$p_{t,h}^{\mathrm{band}} = F_{t,h}(a_t^-) + 1 - F_{t,h}(b_t).$$

Consequently,

$$|p_{t,h}^{\mathrm{band}} - p_t^{\mathrm{band}}| \le |F_{t,h}(a_t^-) - F_t(a_t^-)| + |F_{t,h}(b_t) - F_t(b_t)|.$$

The Kolmogorov bound defining $\Gamma_{t,h}$ also controls left limits, by taking an increasing sequence converging to a_t. Each term is at most $\Gamma_{t,h}$, which proves equation 36. □

B.9 Proof of Corollary B.8

Proof. By the pathwise bound in Theorem B.2,

$$p_{t,h}^{\mathrm{band}} \le p_t^{\mathrm{band}} + 2\Gamma_{t,h} \le \alpha + 2\big(\varepsilon_t + D_t\big) + 2\Gamma_{t,h},$$

which proves equation 43. If the operational and raw intervals coincide, then $p_t^{\mathrm{band}} = p_t^q$, and

$$|p_{t,h}^{\mathrm{band}} - \alpha| \le |p_{t,h}^{\mathrm{band}} - p_t^{\mathrm{band}}| + |p_t^q - \alpha|.$$

Proposition B.3 and Corollary B.7 now give equation 44. □

B.10 Proof of Lemma 3.2

Proof. Write $R = R_{\boldsymbol{\theta}}$ and set $\gamma_{\max} = 2/L_{\min} \le 1$. If $0 \le R \le 1$, then $R^{2/L} \le 1$. If $R > 1$, then $2/L \le \gamma_{\max}$, and concavity of $r \mapsto r^{\gamma_{\max}}$ gives

$$R^{2/L} \le R^{\gamma_{\max}} \le 1 + \gamma_{\max}(R-1).$$

Thus, pointwise,

$$R^{2/L} \le 1 + \gamma_{\max}(R-1)_+.$$

Because $\boldsymbol{y}_U$ has the old-policy marginal, absolute continuity and normalization of the exact sequence ratio imply

$$\mathbb{E}_{\boldsymbol{\theta}_{\text{old}}}[R] = \mathbb{E}_{\boldsymbol{x}\sim p}\left[\sum_{\boldsymbol{y}} \pi_{\boldsymbol{\theta}_{\text{old}}}(\boldsymbol{y} \mid \boldsymbol{x}) \frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{y} \mid \boldsymbol{x})}{\pi_{\boldsymbol{\theta}_{\text{old}}}(\boldsymbol{y} \mid \boldsymbol{x})}\right] = 1.$$

Therefore $\mathbb{E}[(R-1)_+] \le \mathbb{E}[R] = 1$, and

$$\mathbb{E}[W_{\boldsymbol{\theta}}^2] = \mathbb{E}[R^{2/L}] \le 1 + \gamma_{\max} = 1 + \frac{2}{L_{\min}}.$$

□

B.11 Proof of Proposition 3.3

Proof. If $\widehat{A} \ge 0$, then

$$\ell_{a,b}(W_{\boldsymbol{\theta}}, \widehat{A}) = \widehat{A} \min\{W_{\boldsymbol{\theta}}, b\},$$

so the pointwise gap is $\widehat{A}_+(W_{\boldsymbol{\theta}} - b)_+$. If $\widehat{A} < 0$, multiplication reverses the ordering and

$$\ell_{a,b}(W_{\boldsymbol{\theta}}, \widehat{A}) = \widehat{A} \max\{W_{\boldsymbol{\theta}}, a\},$$

so the gap is $\widehat{A}_-(a - W_{\boldsymbol{\theta}})_+$. This proves the nonnegative identity in equation 17.

For the upper tail, Cauchy–Schwarz and Lemma 3.2 give

$$\mathbb{E}[\widehat{A}_+(W_{\boldsymbol{\theta}} - b)_+] \le A_{\max}\mathbb{E}[W_{\boldsymbol{\theta}}\mathbf{1}_{\{W_{\boldsymbol{\theta}} > b\}}] \le A_{\max}\sqrt{\mathbb{E}[W_{\boldsymbol{\theta}}^2]\, p_{\boldsymbol{\theta}}^+} \le A_{\max}\sqrt{\left(1 + \frac{2}{L_{\min}}\right) p_{\boldsymbol{\theta}}^+}.$$

For the lower tail,

$$\mathbb{E}[\widehat{A}_-(a - W_{\boldsymbol{\theta}})_+] \le A_{\max} a\, p_{\boldsymbol{\theta}}^-.$$

Combining the two bounds proves equation 18; using $p_{\boldsymbol{\theta}}^+, p_{\boldsymbol{\theta}}^- \le p_{\boldsymbol{\theta}}$ proves equation 19. □

B.12 Proof of Proposition 3.4

Proof. Because the old policy and L are fixed within the update,

$$\nabla_{\boldsymbol{\theta}} W_{\boldsymbol{\theta}} = W_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\theta}} \log W_{\boldsymbol{\theta}} = W_{\boldsymbol{\theta}} \frac{1}{L} \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\boldsymbol{y} \mid \boldsymbol{x}) = W_{\boldsymbol{\theta}} \psi_{\boldsymbol{\theta}}.$$

Differentiating the unclipped objective gives the identity for g_{G} in equation 21. For $\widehat{A} > 0$, the clipped term is constant only when $W_{\boldsymbol{\theta}} > b$; for $\widehat{A} < 0$, it is constant only when $W_{\boldsymbol{\theta}} < a$. At equality, the specified branch rule selects the unclipped derivative. This gives the identity for the selected direction g_{C} and

$$g_{\text{G}} - g_{\text{C}} = \mathbb{E}[W_{\boldsymbol{\theta}} \widehat{A} \psi_{\boldsymbol{\theta}} \mathbf{1}_{\mathcal{C}_{\boldsymbol{\theta}}}].$$

Using the advantage and score-vector bounds, Cauchy–Schwarz, the event inclusion $\mathbb{P}(\mathcal{C}_{\boldsymbol{\theta}}) \le p_{\boldsymbol{\theta}}$, and Lemma 3.2,

$$\|g_{\text{G}} - g_{\text{C}}\| \le A_{\max} M_\psi \mathbb{E}[W_{\boldsymbol{\theta}} \mathbf{1}_{\mathcal{C}_{\boldsymbol{\theta}}}] \le A_{\max} M_\psi \sqrt{\mathbb{E}[W_{\boldsymbol{\theta}}^2]\mathbb{P}(\mathcal{C}_{\boldsymbol{\theta}})} \le A_{\max} M_\psi \sqrt{\left(1 + \frac{2}{L_{\min}}\right) p_{\boldsymbol{\theta}}}.$$

Finally,

$$\|X_{\text{C}}\|^2 \le A_{\max}^2 M_\psi^2 W_{\boldsymbol{\theta}}^2.$$

Taking expectations and invoking the lemma proves equation 23. □

B.13 Proof of Corollary B.9

Proof. For the group estimator in the corollary, Jensen's inequality gives

$$\|Z_j\|^2 = \left\| \frac{1}{G} \sum_{i=1}^{G} X_{j,i} \right\|^2 \leq \frac{1}{G} \sum_{i=1}^{G} \|X_{j,i}\|^2.$$

Equation equation 23 therefore implies $\mathbb{E}[\|Z_j\|^2 \mid \mathcal{J}] \leq C_{\mathrm{grad}}$; no within-group independence is required. Conditional independence across prompt groups and $\mathbb{E}[Z_j \mid \mathcal{J}] = g_{\mathrm{C}}$ yield

$$\mathbb{E}[\|\widehat{g}_{\mathrm{C}} - g_{\mathrm{C}}\|^2 \mid \mathcal{J}] = \frac{1}{N^2} \sum_{j=1}^{N} \mathbb{E}[\|Z_j - g_{\mathrm{C}}\|^2 \mid \mathcal{J}] \leq \frac{C_{\mathrm{grad}}}{N}.$$

By L_{s}-smoothness,

$$\mathcal{J}_{\mathrm{G}}(\boldsymbol{\theta}^+) \geq \mathcal{J}_{\mathrm{G}}(\boldsymbol{\theta}) + \gamma \langle g_{\mathrm{G}}, \widehat{g}_{\mathrm{C}} \rangle - \frac{L_{\mathrm{s}} \gamma^2}{2} \|\widehat{g}_{\mathrm{C}}\|^2.$$

Taking conditional expectations yields

$$\mathbb{E}[\mathcal{J}_{\mathrm{G}}(\boldsymbol{\theta}^+) - \mathcal{J}_{\mathrm{G}}(\boldsymbol{\theta}) \mid \mathcal{J}] \geq \gamma \langle g_{\mathrm{G}}, g_{\mathrm{C}} \rangle - \frac{L_{\mathrm{s}} \gamma^2}{2} \mathbb{E}[\|\widehat{g}_{\mathrm{C}}\|^2 \mid \mathcal{J}].$$

Proposition 3.4 gives

$$\langle g_{\mathrm{G}}, g_{\mathrm{C}} \rangle \geq \|g_{\mathrm{G}}\| (\|g_{\mathrm{G}}\| - B_{\boldsymbol{\theta}})$$

and $\|g_{\mathrm{C}}\| \leq \|g_{\mathrm{G}}\| + B_{\boldsymbol{\theta}}$. Unbiasedness and the variance bound imply

$$\mathbb{E}[\|\widehat{g}_{\mathrm{C}}\|^2 \mid \mathcal{J}] = \|g_{\mathrm{C}}\|^2 + \mathbb{E}[\|\widehat{g}_{\mathrm{C}} - g_{\mathrm{C}}\|^2 \mid \mathcal{J}] \leq (\|g_{\mathrm{G}}\| + B_{\boldsymbol{\theta}})^2 + \frac{C_{\mathrm{grad}}}{N}.$$

Substitution proves equation 45. Solving for the positive range of γ gives equation 46. □

B.14 Proof of Theorem B.10

Proof. Condition throughout on $\mathcal{G}$. Introduce independent continuous auxiliary variables only for the proof and order the $n+1$ pairs (W_i, U_i) lexicographically. Exchangeability makes the rank of $(W^\star, U^\star)$ uniform on $\{1, \ldots, n+1\}$. The strict event $W^\star < W_{(k_{\mathrm{lo}})}$ implies that this randomized rank is at most k_{lo}, while $W^\star > W_{(k_{\mathrm{hi}})}$ implies that it exceeds k_{hi}. Hence their probabilities are at most $k_{\mathrm{lo}}/(n+1)$ and $(n+1-k_{\mathrm{hi}})/(n+1)$, respectively. The two strict tails are disjoint. The definitions of the floor and ceiling ranks bound their sum by $\alpha_{\mathrm{lo}} + \alpha_{\mathrm{hi}} = \alpha$. Finally, anchoring at 1 only enlarges the interval and can only remove tail events. □

B.15 Proof of Corollary B.11

Proof. The weighted triangle inequality gives $\|\widehat{F}_{\mathcal{B}_t} - \widehat{F}_t\|_\infty \leq \widehat{D}_t$. The strict empirical tails of the buffer at its raw quantile endpoints sum to at most α, and outward rounding and anchoring can only reduce this sum. Comparing the two endpoint CDF values with $\widehat{F}_t$ therefore gives $\widehat{p}_t^{\,\mathrm{band}} \leq \alpha + 2\widehat{D}_t$. Comparing $\widehat{F}_t$ with F_t adds at most $2\xi_t$, proving equation 49.

For each retained step, $\|F_\tau - F_t\|_\infty \leq \|\widehat{F}_\tau - \widehat{F}_t\|_\infty + \xi_\tau + \xi_t$. Averaging yields $D_t \leq \widehat{D}_t + \varepsilon_t + \xi_t$. Finally,

$$F_t(w) - F_t(w^-) \leq \widehat{F}_t(w) - \widehat{F}_t(w^-) + 2\xi_t.$$

Taking the supremum and capping by one proves equation 50. □

B.16 Proof of Corollary B.12

Proof. Theorem 3.6 is a deterministic inequality for each realized buffer and target mixture. Applying it at every step yields the pathwise collection of inequalities; selecting a finite step selects one of them. □

B.17 Proof of Theorem 3.6

Proof. Fix a realized path and write $u = r_t^{\mathcal{G}}$. By equation 27, every $g \in \mathcal{G}_M$ satisfies

$$|\widehat{F}_{\mathcal{B}_t}(g) - \overline{F}_t(g)| \le u, \qquad |\widehat{F}_{\mathcal{B}_t}(g^-) - \overline{F}_t(g^-)| \le u.$$

The weighted block identity in the proof of Theorem B.2 also gives $u \le \varepsilon_t$.

Let $\widetilde{q}_{\mathrm{lo},t}$ and $\widetilde{q}_{\mathrm{hi},t}$ be the outward grid endpoints. If the lower endpoint is the sentinel zero, its strict lower-tail probability is zero. Otherwise it is a grid point no larger than $q_{\mathrm{lo},t}$, and the empirical-inverse definition gives

$$\widehat{F}_{\mathcal{B}_t}(\widetilde{q}_{\mathrm{lo},t}^{-}) \le \alpha_{\mathrm{lo}}.$$

Consequently,

$$F_t(\widetilde{q}_{\mathrm{lo},t}^{-}) \le \alpha_{\mathrm{lo}} + u + D_{t,\mathcal{G}}.$$

If the upper endpoint is $+\infty$, its upper-tail probability is zero. Otherwise it is a grid point no smaller than $q_{\mathrm{hi},t}$, so

$$\widehat{F}_{\mathcal{B}_t}(\widetilde{q}_{\mathrm{hi},t}) \ge 1 - \alpha_{\mathrm{hi}},$$

and hence

$$1 - F_t(\widetilde{q}_{\mathrm{hi},t}) \le \alpha_{\mathrm{hi}} + u + D_{t,\mathcal{G}}.$$

Adding the disjoint strict tails gives $\alpha + 2u + 2D_{t,\mathcal{G}}$. Anchoring further enlarges the interval, and capping a probability by one proves equation 29. □

B.18 Proof of Corollary B.1

Proof. With equal block sizes, $n_t = mK$ and the block weights equal $1/K$. Therefore

$$D_{t,\mathcal{G}} \le D_t \le \frac{1}{K}\sum_{r=1}^{K} \|F_{t-r} - F_t\|_\infty \le \frac{\rho}{K}\sum_{r=1}^{K} r = \frac{\rho(K+1)}{2}.$$

Substitution into Theorem 3.6 proves equation 32. Minimizing its excess term over the available candidate windows gives equation 33; no monotonicity of $r_{t,K}^{\mathcal{G}}$ is used. □

B.19 Proof of Proposition B.5

Proof. Fix a path satisfying the stated conditions. By the definitions of the grid discrepancy and drift, every finite grid point has both strict and non-strict CDF error relative to the step mixture F_t at most

$$\eta_t = r_t^{\mathcal{G}} + D_t.$$

Fix a target p, write $q_t^\star = q_{t,p}^\star$, and choose $\eta_t/\underline{f} < u < r_q$, which is possible by Assumption B.4. Set $z_- = q_t^\star - u$ and $z_+ = q_t^\star + u$. The local CDF condition gives $F_t(z_-) \le p - \underline{f}u$ and $F_t(z_+) \ge p + \underline{f}u$. At the bracketing grid points, monotonicity and the grid-error bound imply $\widehat{F}_{\mathcal{B}_t}(g^-(z_-)) < p$ and $\widehat{F}_{\mathcal{B}_t}(g^+(z_+)) > p$. Hence the empirical quantile and either outward-rounded endpoint lie between these grid points. Their distance from $q_t^\star$ is at most $u + h_{\mathcal{G}}$. Letting $u \downarrow \eta_t/\underline{f}$ yields

$$|\widetilde{q}_{t,p} - q_{t,p}^\star| \le h_{\mathcal{G}} + \frac{r_t^{\mathcal{G}} + D_t}{\underline{f}}. \tag{53}$$

For a window of the preceding K equal-weight CDFs,

$$D_t \le \frac{1}{K}\sum_{r=1}^{K} \|F_{t-r} - F_t\|_\infty \le \frac{1}{K}\sum_{r=1}^{K}\sum_{s=t-r+1}^{t} v_s.$$

After summing over t, each v_s receives total weight at most $K^{-1}\sum_{r=1}^{K} r = (K+1)/2$. Thus

$$\sum_{t=K+1}^{T} D_t \le \frac{K+1}{2} V_T.$$

Summing equation 53 and upper-bounding each realized grid discrepancy by $r^{\mathcal{G}}_{T,K}$ proves equation 37.

The pinball risk is minimized at $q^\star_{t,p}$, and its subgradients in q have absolute value at most one. Its excess is therefore at most $|\widetilde{q}_{t,p} - q^\star_{t,p}|$, which proves the regret statement. Anchoring maps $u \mapsto \min\{u, 1\}$ and $u \mapsto \max\{u, 1\}$ are one-Lipschitz. Finally, apply the exact hinge representation in equation 17 at each collected update and average with the fixed weights $\nu_{t,k}$. This gives

$$|\mathfrak{D}_t(a, b) - \mathfrak{D}_t(a', b')| \le A_{\max}(|a - a'| + |b - b'|).$$

Applying the endpoint bound to both tails and summing proves equation 38. □

B.20 Proof of Proposition B.6

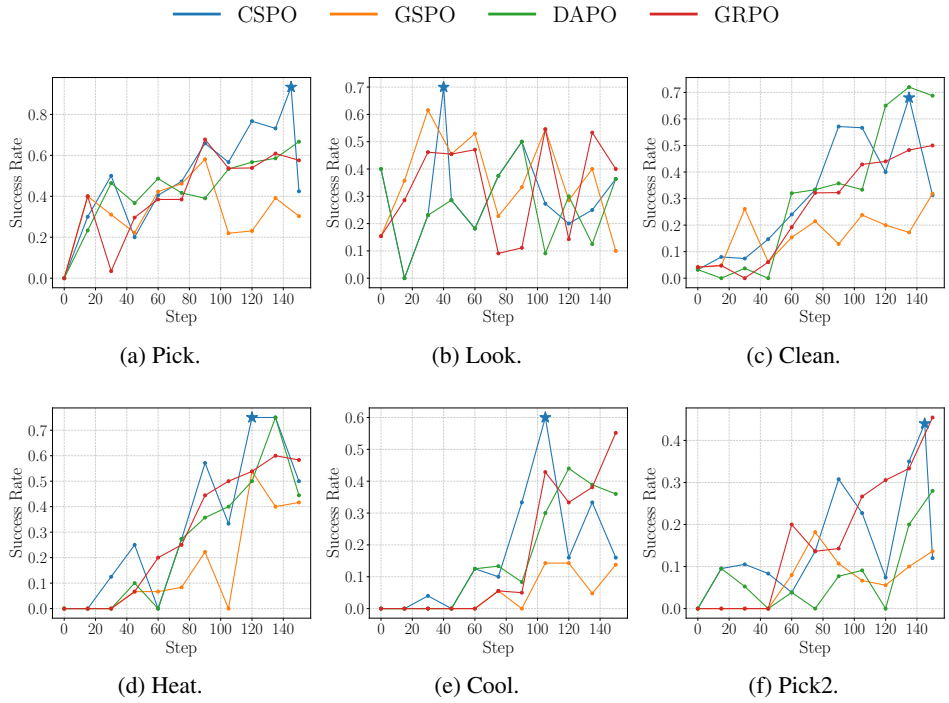
Proof. Let $d_k = \|\boldsymbol{\theta}^{\mathrm{C}}_k - \boldsymbol{\theta}^{\mathrm{G}}_k\|$. Adding and subtracting $g_{\mathrm{G}}(\boldsymbol{\theta}^{\mathrm{C}}_k)$, then using Lipschitzness and the assumed clipping bias, gives

$$d_{k+1} \le (1 + \gamma_k L_g) d_k + \gamma_k B_k.$$

Since $d_0 = 0$, iterating this recursion yields equation 39. The stated choice of B_k follows directly from Proposition 3.4 when its assumptions and the required pathwise tail bound hold at iterate k. □

C SUPPLEMENTARY RESULTS

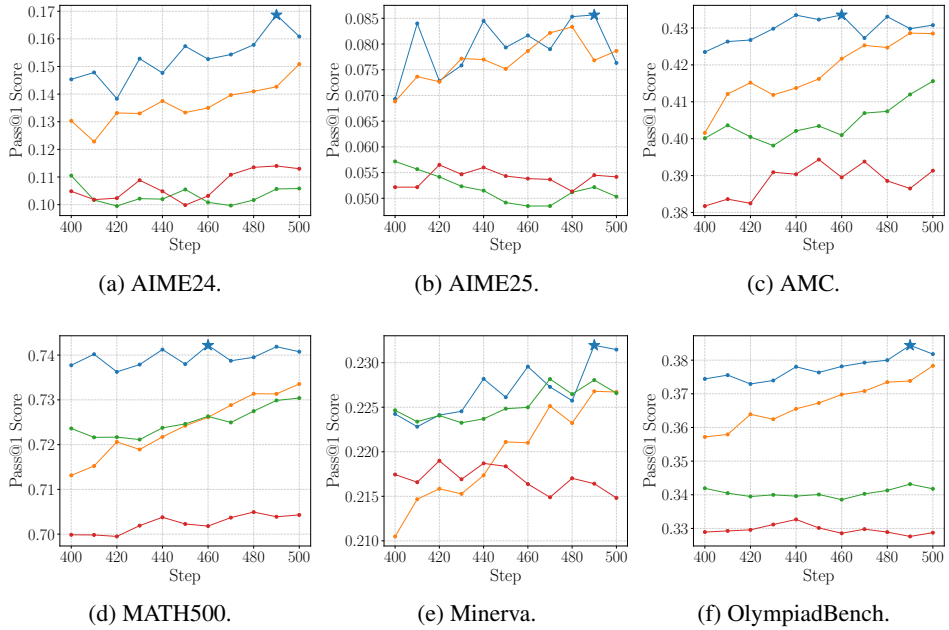
C.1 DETAILED EVALUATION RESULTS



(a) Pick. (b) Look. (c) Clean.

(d) Heat. (e) Cool. (f) Pick2.

Figure 3: Validation results on ALFWorld.



(a) AIME24. (b) AIME25. (c) AMC.

(d) MATH500. (e) Minerva. (f) OlympiadBench.

Figure 4: Pass@1 on various benchmarks.

C.2 Experimental setup

For ALFWorld, we use Qwen3-1.7B and roll out $G = 8$ trajectories per task instance, with a maximum of 50 interaction steps per episode. The maximum prompt and response lengths are 2048 and 512 tokens, respectively. We use a training batch size of 16 task instances and a mini-batch size of 128. All methods are trained for 150 epochs with a learning rate of 1e-6 and a KL penalty coefficient of 0.01.

For DAPO-Math-17k, we use Qwen3-1.7B-Base and roll out $G = 8$ responses per prompt, with maximum prompt and response lengths of 2048 and 8192 tokens, respectively. We use a training batch size of 512 prompts and a mini-batch size of 128, and train all methods for 500 steps with a learning rate of 1e-6. The learning-rate warmup ratio is 0.05 and the weight decay is 0.1.

C.3 Ablation Studies

In the main experiments presented in Section 4.1, CSPO uses $\alpha = 0.1$ and $K = 2$ as its key hyperparameters. These choices were informed by preliminary ablation studies on mathematical reasoning tasks, including MATH (Hendrycks et al., 2021) and GSM8K (Cobbe et al., 2021), following their original train-test splits. As shown in Table 4, performance is relatively stable across a range of K and α, while $K = 2$ and $\alpha = 0.1$ provide a strong and balanced configuration across the two benchmarks. We therefore adopt this setting for the main experiments.

Table 4: Ablation study on window K and target tail budget α. Each column is a (K, α) configuration.

(K, α)	$(2, 0.1)$	$(1, 0.1)$	$(3, 0.1)$	$(4, 0.1)$	$(5, 0.1)$	$(2, 0.05)$	$(2, 0.15)$	$(2, 0.2)$
MATH	65.32	65.82	65.30	66.04	64.68	64.48	65.60	65.16
GSM8K	86.35	85.82	85.44	85.97	85.60	86.20	86.05	85.90
Avg.	75.84	75.82	75.37	76.01	75.14	75.34	75.83	75.53

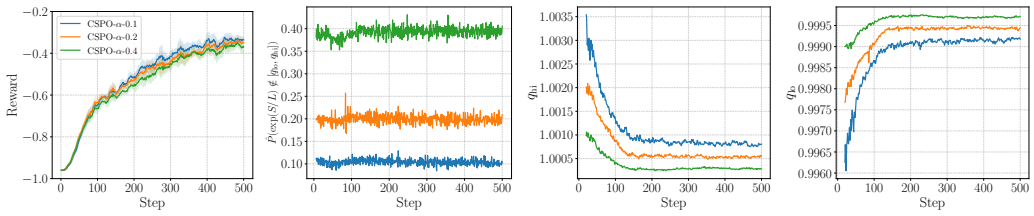
We further study the effect of the miscoverage level α on DAPO-Math-17k. As shown in Table 5, CSPO generally achieves better performance with a smaller α, with $\alpha = 0.1$ obtaining the best average score. A smaller α corresponds to a wider conformal clipping interval and therefore allows more policy updates to remain unclipped. However, α should not be chosen arbitrarily small: as α approaches zero, the clipping interval becomes increasingly loose and CSPO gradually approaches an effectively unclipped policy update, which departs from the trust-region intuition underlying TRPO and PPO and may increase the risk of unstable or collapsed training. Figure 5 shows the rewards, clip fractions, and endpoints for these choices of α. The clip fraction is an empirical training diagnostic; interpreting it as a population tail probability requires accounting for data reuse and distinguishing band exceedance from saturation, which also depends on the advantage sign.

Table 5: Ablations on miscoverage level α.

Method	AIME24	AIME25	AMC	MATH500	Minerva	Oly.	Average
CSPO-α-0.1	16.87	8.57	43.36	74.22	23.19	38.44	34.11
CSPO-α-0.2	14.85	8.58	42.86	74.49	23.02	37.23	33.50
CSPO-α-0.4	13.52	7.40	41.50	72.98	22.47	36.34	32.37

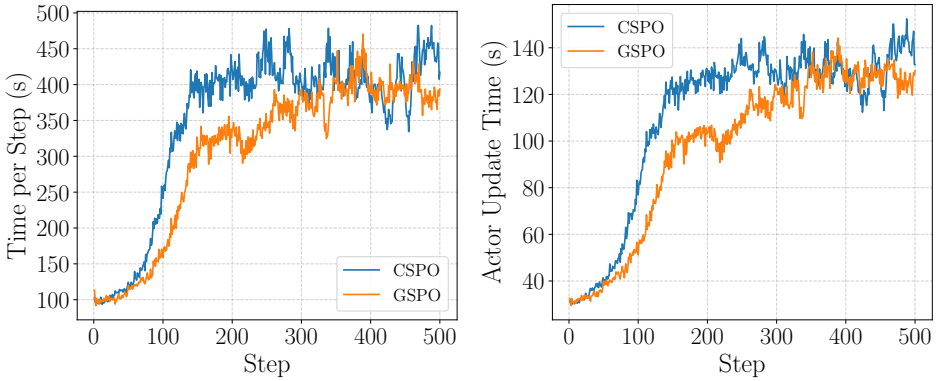
C.4 Additional Training Analysis

Figure 6 shows comparable training-step and actor-update times for CSPO and the baselines, indicating little additional overhead in this setting. Figure 7 reports the PPO KL diagnostic across methods and choices of α. These measurements describe policy changes during training, but do not directly verify the CDF drift assumption in equation 14.



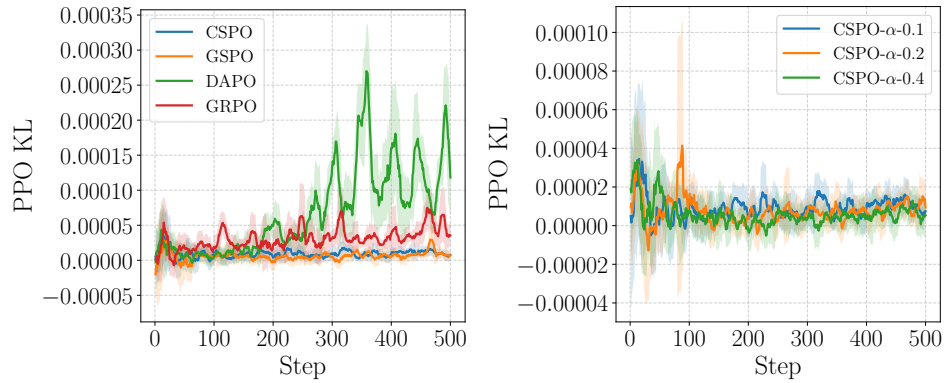
(a) Reward curves. (b) Clip fraction. (c) Dynamics of q_{hi}. (d) Dynamics of q_{lo}.

Figure 5: Ablation studies on DAPO-Math-17k.



(a) Time per training step. (b) Actor update time.

Figure 6: Computational efficiency on DAPO-Math-17k.



(a) PPO KL across methods. (b) PPO KL across different α.

Figure 7: Policy update stability on DAPO-Math-17k.