

Target–ESS Tempered Policy Optimization (TETPO): Smooth Variance Control via Bridge Importance Sampling

Shayan Mohajer Hamidi

Kayhan Behdin

Wenhui Zhu

Aida Rahmattalabi

Jincheng Cao

Jelena Markovic-Voronov

Alborz Geramifard

LinkedIn

SMOHAJER@LINKEDIN.COM

KBEHDIN@LINKEDIN.COM

WENHZHU@LINKEDIN.COM

ARAHMATTALABI@LINKEDIN.COM

JINCAO@LINKEDIN.COM

JMARKOVICVORONOV@LINKEDIN.COM

AGF@LINKEDIN.COM

Abstract

Asynchronous reinforcement-learning post-training of large language models is computationally efficient but inherently off-policy: the learner updates on rollouts generated by a stale policy, while correcting this mismatch requires importance weights that can be extremely heavy-tailed. A small number of trajectories may then dominate an update, collapsing the effective sample size (ESS) and increasing the variance and instability of the resulting gradient estimates. Existing stabilizers apply fixed interventions regardless of the realized severity of weight concentration, either discarding selected samples or reweighting them without adapting the correction strength. We introduce *Target–ESS Tempered Policy Optimization* (TETPO), which replaces each importance weight ρ by ρ^α for a data-adaptive exponent $\alpha \in [0, 1]$: $\alpha = 1$ preserves the original importance weights, whereas $\alpha = 0$ assigns uniform coefficients. For each prompt group, TETPO selects the largest α whose tempered weights satisfy a target ESS fraction κ , applying little tempering to healthy groups and stronger tempering when weight degeneracy is severe. At the population level, the tempered coefficients are exact importance weights for a geometric bridge between the sampler and learner policies. The exponent is efficiently found by one-dimensional bisection, while the resulting weights satisfy deterministic bounds on normalized-weight dispersion and weighted-gradient norm that are independent of policy drift, together with a fixed- α bound on bridge-target discrepancy. Power tempering also introduces no explicit zero-weight mask. On mathematical-reasoning benchmarks under high staleness, TETPO outperforms strong on- and off-policy baselines.

Keywords: off-policy reinforcement learning, large language models, importance sampling, effective sample size, variance reduction

1. Introduction and Motivation

Modern reinforcement-learning (RL) post-training of large language models (LLMs) is dominated by the cost of *rollout generation*: producing several long completions $y \sim \mu(\cdot | x)$ per prompt x , scoring them, and taking policy-gradient steps on the resulting (x, y, R) triples. In strictly synchronous schedules the learner sits idle during sampling. Asynchronous schedules [2, 7, 8, 10] break this coupling: a sampler generates from a policy snapshot several gradient steps in the past while the learner trains on rollouts as they arrive, giving end-to-end speedups of 2–3 $\times$ over synchronous training [8].

The cost is that training becomes *off-policy*. Let $\pi_\theta = \pi_{\theta_t}$ be the learner at step t and $\mu = \pi_{\theta_{t'}}$ a sampler snapshot at step $t' = t - k$. Assuming $\pi_\theta(\cdot | x) \ll \mu(\cdot | x)$ and θ -independent R , importance sampling (IS) yields the gradient identity

$$\nabla_\theta J(\theta) = \mathbb{E}_{x \sim \mathcal{D}, y \sim \mu(\cdot | x)} \left[w(x, y) (R(x, y) - b(x)) \nabla_\theta \log \pi_\theta(y | x) \right], \quad w(x, y) \triangleq \frac{\pi_\theta(y | x)}{\mu(y | x)}, \quad (1)$$

for any y -independent baseline $b(x)$. The group-relative advantages used in the practical surrogate are finite-sample variance-reduction terms; our results below concern the tempered weights conditional on the realised advantages, not unbiasedness of the full clipped update. But $w(x, y) = \prod_{t=1}^T \pi_\theta(y_t | \cdot) / \mu(y_t | \cdot)$ is a product of T per-token ratios, with T in the thousands. Even small per-token shifts compound multiplicatively, so w is extremely heavy-tailed: the *curse of the horizon* [5, 6]. In practice a few trajectories dominate each update, and the *effective sample size*

$$\text{ESS}(\mathbf{w}) \triangleq \frac{(\sum_{i=1}^B w_i)^2}{\sum_{i=1}^B w_i^2} \in [1, B] \quad (2)$$

collapses below the batch size B . When ESS collapses, weight concentration can sharply increase the variance and instability of the resulting gradient estimates, the learner-to-sampler KL spikes, and reward can collapse. The central question is therefore: *how do we control the dispersion of heavy-tailed IS weights without losing so much information that learning stalls?*

Existing stabilizers. Two recent methods take complementary approaches. *M2PO* [13] works per token: with ratio $r_{i,t} = \pi_\theta(y_{i,t} | \cdot) / \mu(y_{i,t} | \cdot)$ and second-moment proxy $M_{i,t}^{(2)} = (\log r_{i,t})^2$, it greedily zeroes the highest- $M^{(2)}$ tokens in a trust region until the mean $M^{(2)}$ falls below a threshold τ_{M_2} , yielding a masked surrogate $\mathcal{L}_{\text{M2PO}} = -\frac{1}{N} \sum_{i,t} m_{i,t} r_{i,t} A_i$ with $m_{i,t} \in \{0, 1\}$. *GEPO* [11] works at the sequence level and never masks: rather than normalizing each likelihood ratio by the sum of the ratios in the batch, it divides the learner sequence-likelihood by a *group-expectation* denominator, i.e. an estimate $\hat{\mathbb{E}}_{\text{grp}}[\pi_\theta]$ of the expected learner likelihood over the group, giving a group-expectation-normalized weight $\hat{w}_i^{\text{GEPO}} = \pi_\theta(y_i) / \hat{\mathbb{E}}_{\text{grp}}[\pi_\theta]$ (computed in log space for stability), so every sample contributes and nothing is discarded.

Failure modes. Each buys stability at a cost that becomes acute as lag k grows. **(i) M2PO collapses under heterogeneous staleness.** The mask is discontinuous in $r_{i,t}$; when prompts in a batch are sampled at different lags, the M^2 statistic is dominated by high-lag tokens and the greedy mask zeroes a large fraction of the trust region, sometimes entirely, so that under mixed-staleness queues M2PO’s trusted-token fraction becomes erratic and the run can crash. In our $k = 32$ runs M2PO peaks early then *declines* on every benchmark. **(ii) GEPO has no variance-control knob.** Its group-expectation normalization is fixed and has no tunable strength, so when a batch is dominated by a few heavy sequences it passes through highly concentrated weights without an explicit ESS constraint; at $k = 32$ GEPO suffers a catastrophic reward drop near step 200.

Our Approach. We want an estimator that (a) **adapts** its correction strength to the groupwise ESS: aggressive when ESS is collapsing, dormant when healthy; (b) is **smooth** in the IS ratios (no discrete masking); (c) provides a **bound on the fixed- α bridge-target discrepancy**; and (d) **cannot collapse**, so weights stay strictly positive even under an adversarial batch which motivate TETPO.

2. Methodology: Target-ESS Tempered Policy Optimization

We follow the off-policy LLM-RL setting of M2PO [13]. A minibatch contains M prompts $x_1, \dots, x_M$ with G completions each; for prompt x_m the completions $y_{m,1}, \dots, y_{m,G}$ are drawn from the recorded behaviour policy $\mu_m(\cdot) := \mu(\cdot \mid x_m)$, and the sequence-level ratio is

$$\rho_{m,i} := \frac{\pi_\theta(y_{m,i} \mid x_m)}{\mu_m(y_{m,i})} = \exp\left(\sum_t \log \frac{\pi_\theta(y_{m,i,t} \mid x_m, y_{m,i,<t})}{\mu_m(y_{m,i,t} \mid x_m, y_{m,i,<t})}\right). \quad (3)$$

Because the bridge normaliser $Z_\alpha(x)$ (Theorem 1) is prompt-dependent, we normalise and select the tempering exponent *within each prompt group*, not across the pooled minibatch (which would implicitly reweight the prompt distribution).

Definition 1 (Groupwise tempered weights and ESS) For $\alpha \in [0, 1]$ and each group m ,

$$\bar{w}_{m,i}(\alpha) := \rho_{m,i}^\alpha, \quad \tilde{w}_{m,i}(\alpha) := \frac{\bar{w}_{m,i}(\alpha)}{\frac{1}{G} \sum_{j=1}^G \bar{w}_{m,j}(\alpha)}, \quad (4)$$

so $\frac{1}{G} \sum_i \tilde{w}_{m,i}(\alpha) = 1$, and the groupwise effective sample size is

$$\text{ESS}_m(\alpha) := \frac{(\sum_i \rho_{m,i}^\alpha)^2}{\sum_i \rho_{m,i}^{2\alpha}} \in [1, G].$$

Definition 2 (Groupwise target-ESS exponent) Given a target ESS fraction $\kappa \in (0, 1]$,

$$\alpha_m^* := \max \left\{ \alpha \in [0, 1] : \text{ESS}_m(\alpha) \geq \kappa G \right\}. \quad (5)$$

Since $\rho_{m,i}^0 = 1$ gives $\text{ESS}_m(0) = G \geq \kappa G$, the feasible set is non-empty; as $\alpha \mapsto \text{ESS}_m(\alpha)$ is continuous and non-increasing (Theorem 2) the maximum exists with $\text{ESS}_m(\alpha_m^*) \geq \kappa G$ for every group. We search all of $[0, 1]$ (no floor $\alpha_{\min}$) by 1-D bisection.

Stop-gradient surrogate. Let $\ell_{m,i}(\theta) := \log \pi_\theta(y_{m,i} \mid x_m)$ and let $\text{sg}[\cdot]$ be the stop-gradient operator. The exponent α_m^* and weights $\tilde{w}_{m,i}(\alpha_m^*)$ are computed from detached ratios. Define the straight-through surrogate ratio

$$r_{m,i}(\theta) := \text{sg}[\tilde{w}_{m,i}(\alpha_m^*)] \exp(\ell_{m,i}(\theta) - \text{sg}[\ell_{m,i}(\theta)]), \quad (6)$$

whose forward value is $\tilde{w}_{m,i}(\alpha_m^*)$ and whose gradient is $\tilde{w}_{m,i}(\alpha_m^*) \nabla_\theta \ell_{m,i}(\theta)$; at $\alpha_m^* = 0$ the method reduces to a uniform-coefficient, GRPO-style score update on the stale samples (not an on-policy update, since the samples still come from μ_m). With group-relative advantage $A_{m,i}$,

$$\mathcal{L}_{\text{TETPO}}(\theta) = -\frac{1}{MG} \sum_{m=1}^M \sum_{i=1}^G \min\left(r_{m,i}(\theta) A_{m,i}, \text{clip}(r_{m,i}(\theta), 1 - \varepsilon_{\text{lo}}, 1 + \varepsilon_{\text{hi}}) A_{m,i}\right), \quad (7)$$

with $A_{m,i}$ and all ESS-selection quantities treated as constants w.r.t. the learner gradient. This is a bridge-weighted PPO surrogate, not the exact gradient of the bridge objective: since $\tilde{\pi}_{\alpha,\theta}$ depends on θ , $\nabla_\theta \mathbb{E}_{\tilde{\pi}_\alpha}[R] = \alpha \mathbb{E}_{\tilde{\pi}_\alpha}[(R - \mathbb{E}_{\tilde{\pi}_\alpha}[R]) \nabla_\theta \log \pi_\theta]$, which (7) does not reproduce.

Two regimes. Per group, (7) interpolates between two ends. *Healthy* ($\text{ESS}_m(1) \geq \kappa G$): $\alpha_m^* = 1$ and the group uses groupwise SNIS coefficients inside the clipped surrogate. *Degenerate*: the bisection returns $\alpha_m^* \in [0, 1)$, tempering toward the uniform-weights limit ($\alpha_m^* = 0$), which is always feasible. No group ever produces a zero tempering weight (Proposition 2), distinguishing TETPO from M2PO’s hard masking.

3. Theoretical Guarantees

TETPO’s tempering enjoys four guarantees, stated and proved in Appendix A; we summarise them here.

Exact bridge target (Thm. 1). Per prompt, the tempered coefficient $\rho^\alpha / Z_\alpha(x)$, with $Z_\alpha(x) = \mathbb{E}_\mu[\rho^\alpha]$, is the exact importance ratio of a geometric-bridge policy $\tilde{\pi}_\alpha \propto \pi_\theta^\alpha \mu^{1-\alpha}$ against μ ; $\alpha = 1$ recovers π_θ and $\alpha = 0$ recovers μ . Because $Z_\alpha(x)$ is prompt-dependent it cancels *within* a group but not across prompts, which is why normalisation is groupwise. This characterises the population target, not the finite-sample estimator (which is biased; see below).

Well-posed selection (Thm. 2). $\alpha \mapsto \text{ESS}_m(\alpha)$ is continuous and non-increasing with $\text{ESS}_m(0) = G$, so for every $\kappa \in (0, 1]$ the target-ESS exponent α_m^* exists (uniquely in the degenerate case) and is found by bisection on $[0, 1]$; there is no need for a floor $\alpha_{\min}$.

ESS-controlled dispersion (Thm. 3). The self-normalised weights obey $\frac{1}{G} \sum_i (\tilde{w}_{m,i}(\alpha_m^*) - 1)^2 = G / \text{ESS}_m(\alpha_m^*) - 1 \leq 1/\kappa - 1$, independent of the π_θ -to- μ divergence, giving a weighted-gradient-norm bound $\|\hat{g}_m\|^2 \leq \frac{1}{\kappa} \cdot \frac{1}{G} \sum_i \|g_{m,i}\|^2$. This is a deterministic realised-batch bound, not a bound on the estimator’s stochastic variance; in contrast, ordinary IS has $\text{Var}_\mu[\rho] = e^{D_2(\pi_\theta \parallel \mu)} - 1$, which grows with the drift (D_2 the order-two Rényi divergence).

Bounded discrepancy (Thm. 4). For fixed α and $|R| \leq R_{\max}$, the bridge objective differs from the on-policy objective by at most $R_{\max} \sqrt{\chi^2(\tilde{\pi}_\alpha \parallel \pi_\theta)}$. As $\chi^2(\tilde{\pi}_\alpha \parallel \pi_\theta)$ is non-increasing in α (Prop. 1) while dispersion is non-decreasing, α_m^* is the least-tempered exponent meeting the ESS target, i.e. it minimises this χ^2 bound among feasible exponents. The finite-sample and adaptation bias of the self-normalised estimator is not characterised.

No masking (Prop. 2). All tempered weights are strictly positive, so tempering adds no explicit zero-weight mask, unlike M2PO, whose greedy mask can zero an entire trust region (the instability observed at high staleness). Positivity alone does not force a nonzero final PPO gradient when the advantage is zero or the clip is active.

Complexity. Beyond a vanilla GRPO step, TETPO adds only $O(n_{\text{iter}} \cdot MG)$ work: the $\log \rho_{m,i}$ are already available, and the 24-iteration per-group bisections plus $O(G)$ self-normalisation are negligible next to the per-token forward/backward passes. Table 2 (Appendix A) contrasts TETPO with prior stabilizers.

4. Experiments

We evaluate TETPO against GRPO [9], GSPO [12], GEPO [11], and M2PO [13] on math-reasoning benchmarks under high staleness ($k = 32$). Table 1 reports avg@16 accuracy (%): TETPO attains the best average on both models, improving over the strongest baseline by +2.1 points on

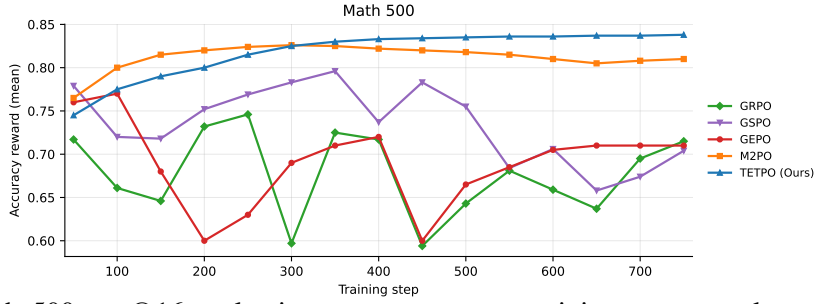


Figure 1: Math 500 avg@16 evaluation accuracy versus training step at staleness $k = 32$, for GRPO, GSPO, GEPO, M2PO, and TETPO (ours). GEPO collapses intermittently and M2PO peaks then declines, whereas TETPO rises smoothly and remains stable.

Table 1: Accuracy (%) across mathematical-reasoning benchmarks, comparing GRPO, GSPO, GEPO, M2PO, and TETPO. Values are avg@16 accuracy under stochastic sampling, reported as mean $\pm$ std over three seeds. Within each model block, the **best** result per column is in bold and the second best is underlined.

Method	AIME24	AIME25	AMC23	AMC24	Math500	Avg.
<i>Qwen3-Base-1.7B</i>						
GRPO [9]	15.6 $\pm$ 1.1	17.5 $\pm$ 1.4	42.1 $\pm$ 0.8	21.2 $\pm$ 1.0	74.6 $\pm$ 0.5	34.2 $\pm$ 0.6
GSPO [12]	23.0 $\pm$ 0.9	18.7 $\pm$ 1.2	48.0 $\pm$ 0.7	25.6 $\pm$ 0.9	79.6 $\pm$ 0.4	39.0 $\pm$ 0.5
GEPO [11]	19.5 $\pm$ 1.6	17.9 $\pm$ 1.7	43.4 $\pm$ 1.2	23.1 $\pm$ 1.3	77.1 $\pm$ 0.8	36.2 $\pm$ 0.9
M2PO [13]	<u>24.1</u> $\pm$ 0.7	<u>24.2</u> $\pm$ 1.0	<u>53.1</u> $\pm$ 0.6	<u>27.5</u> $\pm$ 0.8	<u>82.5</u> $\pm$ 0.4	<u>42.3</u> $\pm$ 0.5
TETPO (Ours)	26.6 $\pm$ 0.6	24.6 $\pm$ 0.8	54.3 $\pm$ 0.5	32.8 $\pm$ 0.7	83.8 $\pm$ 0.3	44.4 $\pm$ 0.4
<i>Llama-3.2-3B-Instruct</i>						
GRPO [9]	1.5 $\pm$ 0.7	0.0 $\pm$ 0.0	15.5 $\pm$ 1.1	5.8 $\pm$ 0.9	40.2 $\pm$ 0.6	12.6 $\pm$ 0.5
GSPO [12]	5.4 $\pm$ 0.8	<u>0.6</u> $\pm$ 0.5	20.3 $\pm$ 0.9	9.8 $\pm$ 1.0	44.6 $\pm$ 0.5	16.1 $\pm$ 0.4
GEPO [11]	2.8 $\pm$ 0.9	0.0 $\pm$ 0.0	22.1 $\pm$ 1.3	<u>10.2</u> $\pm$ 1.1	<u>45.9</u> $\pm$ 0.7	<u>16.2</u> $\pm$ 0.6
M2PO [13]	<u>6.6</u> $\pm$ 0.6	0.3 $\pm$ 0.4	19.1 $\pm$ 0.8	8.7 $\pm$ 0.7	43.5 $\pm$ 0.5	15.6 $\pm$ 0.4
TETPO (Ours)	8.2 $\pm$ 0.7	1.6 $\pm$ 0.6	<u>21.4</u> $\pm$ 0.6	14.9 $\pm$ 0.9	46.2 $\pm$ 0.4	18.5 $\pm$ 0.3

Qwen3-Base-1.7B and +2.3 on Llama-3.2-3B-Instruct, and, consistent with the no-collapse guarantee, avoids the mid-training reward drop seen for GEPO and M2PO. Figure 1 shows the Math 500 trajectories at $k = 32$: GEPO collapses sharply (steps 200, 450) as low-ESS batches pass through highly concentrated weights, M2PO peaks then declines as its mask grows aggressive, and the on-policy baselines GRPO/GSPO plateau lower, whereas TETPO rises smoothly to the highest reward. Full details are in Appendix B.

5. Limitations

Tempering rescales all group weights uniformly; Theorem 4 bounds the fixed- α target discrepancy, but the finite-sample and adaptation bias of the estimator is not yet characterised. TETPO also uses sequence-level ratios; combining it with a per-token mask is future work.

References

- [1] Rui Cai, Wenhui Zhu, Xiwen Chen, Jincheng Cao, Han Yu, Shayan Mohajer Hamidi, Zelin He, Qiyao Ma, Daiwei Chen, Xuanzhao Dong, et al. Smaller models, better rejects: Preference distillation scaling. *arXiv preprint arXiv:2609.38987*, 2026.
- [2] Wei Fu, Jiaxuan Gao, Xujie Shen, Chen Zhu, Zhiyu Mei, Chuyi He, Shusheng Xu, Guo Wei, Jun Mei, Jiakai Wang, Tingxun Yang, Binxing Yuan, and Yi Wu. AReaL: A large-scale asynchronous reinforcement learning system for language reasoning. *arXiv preprint arXiv:2505.24298*, 2025.
- [3] Shayan Mohajer Hamidi, Yize Cheng, Yuanda Xu, Zhengze Zhou, and Alborz Geramifard. Hierarchical reinforcement learning with stable temporal abstraction for language model agents. In *SLM-Agents: 1st NeurIPS Workshop on SLMs for Agentic Systems*, 2026. URL <https://openreview.net/forum?id=XYhgL4FluZ>.
- [4] Shayan Mohajer Hamidi, Linfeng Ye, and Konstantinos N. Plataniotis. Dreamphase: Offline imagination and uncertainty-guided planning for large-language-model agents. In *The Fourteenth International Conference on Learning Representations*, 2026. URL <https://openreview.net/forum?id=81PJ2KPnmK>.
- [5] Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. *Advances in neural information processing systems*, 31, 2018.
- [6] Yao Liu, Pierre-Luc Bacon, and Emma Brunskill. Understanding the curse of horizon in off-policy evaluation via conditional importance sampling. In *International Conference on Machine Learning*, pages 6184–6193. PMLR, 2020.
- [7] Michael Noukhovitch, Shengyi Huang, Sophie Xhonneux, Arian Hosseini, Rishabh Agarwal, and Aaron Courville. Asynchronous RLHF: Faster and more efficient off-policy RL for language models. *arXiv preprint arXiv:2410.18252*, 2024.
- [8] Alexandre Piché, Ehsan Kamaloo, Rafael Pardinás, and Dzmitry Bahdanau. PipelineRL: Faster on-policy reinforcement learning for long sequence generation. *arXiv preprint arXiv:2509.19128*, 2025.
- [9] 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.
- [10] Bo Wu, Sid Wang, Yu Tang, Jay Ding, Eric Helenowski, Liang Tan, Tao Xu, Thamme Gowda, Zhuokai Chen, Chen Zhu, Xin Tang, Yongqiang Qian, Bin Zhu, and Rui Hou. LlamaRL: A distributed asynchronous reinforcement learning framework for efficient large-scale LLM training. *arXiv preprint arXiv:2505.24034*, 2025.
- [11] Han Zhang, RuibinZheng, ZEXUAN YI, Zhuo Zhang, Hanyang Peng, Hui Wang, Jiayin Qi, Binxing Fang, Ruifeng Xu, and Yue Yu. GEPO: Group expectation policy optimization

for stable heterogeneous reinforcement learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL <https://openreview.net/forum?id=movNwgQtDt>.

- [12] Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.
- [13] Haizhong Zheng, Jiawei Zhao, and Beidi Chen. Prosperity before collapse: How far can off-policy RL reach with stale data on LLMs? In *The Fourteenth International Conference on Learning Representations*, 2026. URL <https://openreview.net/forum?id=IIgl5MWelz>.

Appendix A. Theoretical results and proofs

Table 2: TETPO is the only method that (i) operates smoothly, (ii) achieves a divergence-free (ESS-controlled) weight-dispersion bound, and (iii) provably adds no explicit zero-weight mask. Second-moment entries use the order-two Rényi divergence D_2 ($\chi^2 = e^{D_2} - 1$).

Method	Granularity	Smoothness	Var./disp. bound	Mask-collapse risk
GSPO [12]	sequence	PPO clip	e^{D_2}	low
GEPO [11]	group	group-exp.	$e^{D_2} - C$	low
M2PO [13]	token	hard mask	$R^2 M_2$	high
SNIS+ESS	sequence	hard truncate	$1/\beta - 1$	low
TETPO (ours)	sequence	smooth	$\boxed{1/\kappa - 1} (disp.)$	none (Prop. 2)

Theorem 1 (Geometric-bridge importance identity) *Fix a prompt x with $\pi_\theta(\cdot | x), \mu(\cdot | x) > 0$ on a common support $\mathcal{Y}_x$. For $\alpha \in [0, 1]$ let $Z_\alpha(x) := \mathbb{E}_{y \sim \mu(\cdot | x)}[\rho(y | x)^\alpha]$ and define the geometric-bridge policy $\tilde{\pi}_\alpha(y | x) := \pi_\theta(y | x)^\alpha \mu(y | x)^{1-\alpha} / Z_\alpha(x)$. Then $\tilde{\pi}_\alpha(y | x) / \mu(y | x) = \rho(y | x)^\alpha / Z_\alpha(x)$, so for every integrable h ,*

$$\mathbb{E}_{y \sim \tilde{\pi}_\alpha(\cdot | x)}[h(y)] = \frac{\mathbb{E}_{y \sim \mu(\cdot | x)}[\rho(y | x)^\alpha h(y)]}{\mathbb{E}_{y \sim \mu(\cdot | x)}[\rho(y | x)^\alpha]}, \quad (8)$$

with endpoints $\tilde{\pi}_0 = \mu$ and $\tilde{\pi}_1 = \pi_\theta$.

Proof of Theorem 1 By the definition of $\tilde{\pi}_\alpha$,

$$\frac{\tilde{\pi}_\alpha(y | x)}{\mu(y | x)} = \frac{\pi_\theta(y | x)^\alpha \mu(y | x)^{1-\alpha}}{Z_\alpha(x) \mu(y | x)} = \frac{1}{Z_\alpha(x)} \left(\frac{\pi_\theta(y | x)}{\mu(y | x)} \right)^\alpha = \frac{\rho(y | x)^\alpha}{Z_\alpha(x)}. \quad (9)$$

Hence, for any integrable h ,

$$\mathbb{E}_{y \sim \tilde{\pi}_\alpha(\cdot | x)}[h(y)] = \sum_{y \in \mathcal{Y}_x} \tilde{\pi}_\alpha(y | x) h(y) = \sum_{y \in \mathcal{Y}_x} \mu(y | x) \frac{\rho(y | x)^\alpha}{Z_\alpha(x)} h(y) = \frac{\mathbb{E}_{y \sim \mu(\cdot | x)}[\rho(y | x)^\alpha h(y)]}{Z_\alpha(x)}. \quad (10)$$

Taking $h \equiv 1$ gives $Z_\alpha(x) = \mathbb{E}_\mu[\rho^\alpha]$, so

$$\mathbb{E}_{\tilde{\pi}_\alpha}[h] = \frac{\mathbb{E}_\mu[\rho^\alpha h]}{\mathbb{E}_\mu[\rho^\alpha]}, \quad (11)$$

which is the stated ratio-of-expectations identity. The endpoints follow from $\tilde{\pi}_0 \propto \mu$ and $\tilde{\pi}_1 \propto \pi_\theta$. Finally, $Z_\alpha(x)$ depends on x , so it cancels between numerator and denominator only within a fixed prompt and not under pooled cross-prompt normalisation. ■

Theorem 2 (Existence, monotonicity, uniqueness, bisection) *Fix a group with $\rho_1, \dots, \rho_G > 0$ and set $f(\alpha) := \text{ESS}_m(\alpha)$ on $[0, 1]$. Then f is continuous, $f(0) = G$, $f(1) = \text{ESS}(\rho)$, and f is non-increasing (strictly decreasing on $(0, 1]$ when the ρ_i are not all equal). Hence for every $\kappa \in (0, 1]$ the feasible set $\{\alpha : f(\alpha) \geq \kappa G\}$ is a non-empty closed interval $[0, \alpha_m^*]$: if $f(1) \geq \kappa G$ then $\alpha_m^* = 1$; if $\kappa < 1$ and $f(1) < \kappa G$ then $\alpha_m^* \in (0, 1)$ is the unique root of $f(\alpha) = \kappa G$; and if $\kappa = 1$ with non-constant ratios then $\alpha_m^* = 0$. Bisection on $[0, 1]$ brackets α_m^* to width $2^{-n_{\text{iter}}}$.*

Proof of Theorem 2 Write $z_i := \log \rho_i$ and

$$S(a) := \sum_{i=1}^G e^{az_i}, \quad f(a) = \frac{S(a)^2}{S(2a)}. \quad (12)$$

Then f is continuous on $[0, 1]$, and

$$f(0) = \frac{G^2}{G} = G, \quad f(1) = \frac{(\sum_i \rho_i)^2}{\sum_i \rho_i^2} = \text{ESS}(\rho). \quad (13)$$

Introduce the exponentially tilted law $p_a(i) := e^{az_i}/S(a)$. Since $\frac{d}{da} \log S(a) = \mathbb{E}_{p_a}[z]$,

$$\frac{d}{da} \log f(a) = 2 \frac{d}{da} \log S(a) - \frac{d}{da} \log S(2a) = 2(\mathbb{E}_{p_a}[z] - \mathbb{E}_{p_{2a}}[z]), \quad (14)$$

and, moreover,

$$\frac{d}{da} \mathbb{E}_{p_a}[z] = \text{Var}_{p_a}(z) \geq 0. \quad (15)$$

Thus $a \mapsto \mathbb{E}_{p_a}[z]$ is non-decreasing, so $\frac{d}{da} \log f(a) \leq 0$ and f is non-increasing; it is strictly decreasing on $(0, 1]$ whenever the z_i are not all equal, since then $\text{Var}_{p_a}(z) > 0$ for finite a . Because $f(0) = G \geq \kappa G$, the feasible set $\{a \in [0, 1] : f(a) \geq \kappa G\}$ is non-empty; by continuity it is closed, and by monotonicity it is an interval $[0, \alpha_m^*]$. The stated cases follow: (i) if $f(1) \geq \kappa G$ then $\alpha_m^* = 1$; (ii) if $\kappa < 1$ and $f(1) < \kappa G$, the intermediate value theorem yields a unique root of $f(\alpha) = \kappa G$ in $(0, 1)$; (iii) if $\kappa = 1$ with non-constant ratios, $f(\alpha) < G$ for all $\alpha > 0$, forcing $\alpha_m^* = 0$; (iv) if all ratios are equal, $f \equiv G$ and $\alpha_m^* = 1$. Bisection halves the bracket at each iteration. $\blacksquare$

Theorem 3 (ESS-controlled dispersion and weighted-gradient norm) *For each group m the self-normalised weights $\tilde{w}_{m,i}(\alpha_m^*)$ (empirical mean 1) satisfy the deterministic dispersion bound*

$$\frac{1}{G} \sum_{i=1}^G (\tilde{w}_{m,i}(\alpha_m^*) - 1)^2 = \frac{G}{\text{ESS}_m(\alpha_m^*)} - 1 \leq \frac{1}{\kappa} - 1, \quad (16)$$

independent of any π_θ -to- μ divergence. Consequently, for arbitrary vectors $g_{m,i}$, the weighted group mean $\hat{g}_m := \frac{1}{G} \sum_i \tilde{w}_{m,i}(\alpha_m^*) g_{m,i}$ obeys $\|\hat{g}_m\|^2 \leq \frac{1}{\kappa} \cdot \frac{1}{G} \sum_i \|g_{m,i}\|^2$, and the minibatch mean $\hat{g} := \frac{1}{M} \sum_m \hat{g}_m$ obeys $\|\hat{g}\|^2 \leq \frac{1}{\kappa M G} \sum_{m,i} \|g_{m,i}\|^2$.

Proof of Theorem 3 Fix a group m and abbreviate $\tilde{w}_i := \tilde{w}_{m,i}(\alpha_m^*)$. Self-normalisation gives $\frac{1}{G} \sum_i \tilde{w}_i = 1$, i.e. $\sum_i \tilde{w}_i = G$. As ESS is invariant to a common positive rescaling of the weights,

$$\text{ESS}_m(\alpha_m^*) = \frac{(\sum_i \tilde{w}_i)^2}{\sum_i \tilde{w}_i^2} = \frac{G^2}{\sum_i \tilde{w}_i^2} \implies \frac{1}{G} \sum_{i=1}^G \tilde{w}_i^2 = \frac{G}{\text{ESS}_m(\alpha_m^*)}. \quad (17)$$

Therefore, using $\frac{1}{G} \sum_i \tilde{w}_i = 1$ and $\text{ESS}_m(\alpha_m^*) \geq \kappa G$,

$$\frac{1}{G} \sum_{i=1}^G (\tilde{w}_i - 1)^2 = \frac{1}{G} \sum_i \tilde{w}_i^2 - 1 = \frac{G}{\text{ESS}_m(\alpha_m^*)} - 1 \leq \frac{1}{\kappa} - 1. \quad (18)$$

For the gradient bound, the Cauchy–Schwarz inequality gives

$$\left\| \frac{1}{G} \sum_{i=1}^G \tilde{w}_i g_i \right\|^2 \leq \left(\frac{1}{G} \sum_i \tilde{w}_i^2 \right) \left(\frac{1}{G} \sum_i \|g_i\|^2 \right) \leq \frac{1}{\kappa} \cdot \frac{1}{G} \sum_{i=1}^G \|g_i\|^2. \quad (19)$$

Averaging over groups and applying Jensen’s inequality,

$$\|\hat{g}\|^2 = \left\| \frac{1}{M} \sum_{m=1}^M \hat{g}_m \right\|^2 \leq \frac{1}{M} \sum_{m=1}^M \|\hat{g}_m\|^2 \leq \frac{1}{\kappa M G} \sum_{m=1}^M \sum_{i=1}^G \|g_{m,i}\|^2. \quad (20)$$

■

Theorem 4 (Fixed- α bridge-target discrepancy) *Fix a prompt x and $\alpha \in [0, 1]$ with $\tilde{\pi}_\alpha(\cdot | x) \ll \pi_\theta(\cdot | x)$, and suppose $|R| \leq R_{\max}$. Then*

$$\left| \mathbb{E}_{\tilde{\pi}_\alpha(\cdot | x)}[R] - \mathbb{E}_{\pi_\theta(\cdot | x)}[R] \right| \leq R_{\max} \sqrt{\chi^2(\tilde{\pi}_\alpha(\cdot | x) \| \pi_\theta(\cdot | x))}, \quad (21)$$

where $\chi^2(p \| q) = \mathbb{E}_q[(p/q - 1)^2]$ is the Pearson divergence.

Proof of Theorem 4 Let $r_\alpha := \tilde{\pi}_\alpha(\cdot | x) / \pi_\theta(\cdot | x)$, which is well defined because $\tilde{\pi}_\alpha \ll \pi_\theta$, and note $\mathbb{E}_{\pi_\theta}[r_\alpha] = 1$. Then

$$\mathbb{E}_{\tilde{\pi}_\alpha}[R] - \mathbb{E}_{\pi_\theta}[R] = \mathbb{E}_{\pi_\theta}[r_\alpha R] - \mathbb{E}_{\pi_\theta}[R] = \mathbb{E}_{\pi_\theta}[(r_\alpha - 1)R]. \quad (22)$$

By the Cauchy–Schwarz inequality and $|R| \leq R_{\max}$,

$$\left| \mathbb{E}_{\pi_\theta}[(r_\alpha - 1)R] \right| \leq \left(\mathbb{E}_{\pi_\theta}[(r_\alpha - 1)^2] \right)^{1/2} \left(\mathbb{E}_{\pi_\theta}[R^2] \right)^{1/2} \leq R_{\max} \sqrt{\chi^2(\tilde{\pi}_\alpha \| \pi_\theta)}, \quad (23)$$

where the last step uses $\mathbb{E}_{\pi_\theta}[(r_\alpha - 1)^2] = \chi^2(\tilde{\pi}_\alpha \| \pi_\theta)$. The bound concerns the population objective $\mathbb{E}_{\tilde{\pi}_\alpha}[R]$ only; it does not assert unbiasedness of the finite-sample self-normalised estimator, nor account for the data-adaptive selection of α_m^* . ■

Proposition 1 (Monotonicity of the bridge discrepancy) *Fix a prompt x with $\mu(\cdot | x)$ and $\pi_\theta(\cdot | x)$ mutually absolutely continuous. Let $X := \log(\mu(Y | x) / \pi_\theta(Y | x))$ with $Y \sim \pi_\theta(\cdot | x)$, and suppose $M(t) := \mathbb{E}_{\pi_\theta}[e^{tX}] < \infty$ for $t \in [0, 2]$. Then $\alpha \mapsto \chi^2(\tilde{\pi}_\alpha \| \pi_\theta)$ is non-increasing on $[0, 1]$, and strictly decreasing on $[0, 1)$ unless $\mu = \pi_\theta$ almost surely.*

Proof Set $t := 1 - \alpha \in [0, 1]$ and $K(t) := \log M(t)$. The bridge density relative to π_θ is

$$\frac{\tilde{\pi}_\alpha(Y | x)}{\pi_\theta(Y | x)} = \frac{e^{tX(Y)}}{M(t)}, \quad (24)$$

so

$$1 + \chi^2(\tilde{\pi}_\alpha \| \pi_\theta) = \mathbb{E}_{\pi_\theta} \left[\left(\frac{e^{tX}}{M(t)} \right)^2 \right] = \frac{M(2t)}{M(t)^2} = \exp(K(2t) - 2K(t)). \quad (25)$$

Differentiating,

$$\frac{d}{dt} \log(1 + \chi^2(\tilde{\pi}_\alpha \| \pi_\theta)) = 2K'(2t) - 2K'(t) \geq 0, \quad (26)$$

because K is a cumulant generating function and hence convex, so K' is non-decreasing. Thus χ^2 is non-decreasing in $t = 1 - \alpha$, i.e. non-increasing in α , and strictly so when X is non-degenerate. ■

Proposition 2 (No hard masking by power tempering) *For every $\alpha \in [0, 1]$ and strictly positive $\rho_1, \dots, \rho_G$, the normalised tempered weights satisfy $\tilde{w}_{m,i}(\alpha) > 0$. Thus power tempering never assigns an exactly zero coefficient to a sampled trajectory. (Strict positivity does not by itself imply a nonzero final PPO gradient: a sample still contributes zero when its advantage is zero or the clip is active.)*

Proof of Proposition 2 Fix $\alpha \in [0, 1]$. If $\alpha > 0$, then $\rho_{m,i} > 0$ implies $\rho_{m,i}^\alpha > 0$; if $\alpha = 0$, then $\rho_{m,i}^0 = 1 > 0$. In either case the group denominator $\frac{1}{G} \sum_j \rho_{m,j}^\alpha > 0$, so

$$\tilde{w}_{m,i}(\alpha) = \frac{\rho_{m,i}^\alpha}{\frac{1}{G} \sum_{j=1}^G \rho_{m,j}^\alpha} > 0. \quad (27)$$

■

Appendix B. Experimental Details and Reproducibility

Models and data. We post-train *Qwen3-Base-1.7B* (and, in the second block, *Llama-3.2-3B-Instruct*) with RL. Training prompts are the SimpleRL-Zoo competition-math split `SimpleLR/simplelr_qwen_level3to5` (difficulty levels 3–5), loaded from the released `train.parquet` and filtered to prompts of at most 768 tokens; we use the entire filtered split without subsampling. Evaluation uses five *separate* public benchmarks: MATH-500 (HuggingFaceH4/MATH-500), AMC23 (HanlardResearch/amc23), AMC24 (rawsh/2024_AMC12), AIME24 (HanlardResearch/aime_2024), and AIME25 (HanlardResearch/aime_2025), which come from sources disjoint from the training corpus, so evaluation items are excluded from training by construction; we did not run an additional n -gram decontamination pass. The reward is binary correctness computed by `math-verify (v0.5.2)`: the gold answer is wrapped in `\boxed{\}` and parsed, the completion is parsed with `boxed-answer` priority, and the reward is `1\{verify(answer, gold)\} \in \{0, 1\}`, with no learned reward model. Rewards are not scaled (`scale_rewards=False`); advantages are the within-group mean-subtracted rewards (GRPO-style).

Optimisation hyperparameters. All methods share: optimiser AdamW, learning rate 1×10^{-6} with cosine schedule and warmup ratio 0.05, gradient clipping 1.0, KL coefficient $\beta = 0$, and 3 epochs with evaluation/checkpointing every 64 steps. Rollouts use $G = 8$ completions per prompt with $M = 64$ prompts per optimisation step (effective batch $B = MG = 512$ sequences; realised as per-device batch $8 \times$ gradient accumulation 8 across 4 GPUs), max prompt length 768 and max completion length 4096. Decoding for rollouts uses temperature 0.6, top- p 0.95, top- k 20, min- p 0. The PPO clip bounds are $\varepsilon_{lo} = \varepsilon_{hi} = 0.2$. For TETPO we use target ESS fraction $\kappa = 0.85$, floor $\alpha_{min} = 0$, and a 24-iteration bisection over $[0, 1]$ ($\sim 10^{-7}$ precision).

Staleness protocol. We use a PipelineRL-style schedule (`uniform_stale_k`): the sampler serves from a pool of recent learner snapshots, and each rollout’s policy lag is drawn independently as $\ell \sim \text{Uniform}\{1, \dots, k\}$ with $k = 32$. Thus k is the *maximum* lag and the realised lag is random (not fixed or averaged); the snapshot pool holds up to 256 checkpoints. This produces the heterogeneous staleness that TETPO’s per-group adaptation targets.

Importance ratios under truncated decoding. Our rollout sampler applies temperature scaling followed by hard top- k /top- p truncation. Let $\mathcal{S}_\mu(h)$ denote the token support retained by this decoding rule at history h . The resulting behaviour distribution is

$$\mu^{\mathcal{S}}(a \mid h) = \frac{\mu^{\text{raw}}(a \mid h) \mathbf{1}\{a \in \mathcal{S}_\mu(h)\}}{\sum_{a' \in \mathcal{S}_\mu(h)} \mu^{\text{raw}}(a' \mid h)},$$

where μ^{raw} denotes the temperature-scaled distribution before top- k /top- p truncation. Because $\mu^{\mathcal{S}}(a \mid h) = 0$ outside $\mathcal{S}_\mu(h)$ while the learner generally assigns positive probability there, trajectories generated by $\mu^{\mathcal{S}}$ cannot provide an exact importance-sampling estimator for the unrestricted learner policy. An exact change-of-measure identity is obtained instead for the learner restricted to the same behaviour support,

$$\pi_\theta^{\mathcal{S}}(a \mid h) = \frac{\pi_\theta^{\text{raw}}(a \mid h) \mathbf{1}\{a \in \mathcal{S}_\mu(h)\}}{\sum_{a' \in \mathcal{S}_\mu(h)} \pi_\theta^{\text{raw}}(a' \mid h)},$$

for which

$$\rho^{\mathcal{S}}(y \mid x) = \prod_t \frac{\pi_\theta^{\mathcal{S}}(y_t \mid x, y_{<t})}{\mu^{\mathcal{S}}(y_t \mid x, y_{<t})}$$

satisfies $\pi_\theta^{\mathcal{S}} \ll \mu^{\mathcal{S}}$ and $\mathbb{E}_{y \sim \mu^{\mathcal{S}}}[\rho^{\mathcal{S}}(y \mid x)] = 1$.

In the experiments reported here, however, we do not retain $\mathcal{S}_\mu(h)$ or the corresponding truncation normalisers. We record the log-probability of the sampled token under the pre-truncation, temperature-scaled behaviour distribution and form the sequence-level proxy

$$\hat{\rho}(y \mid x) = \prod_t \frac{\pi_\theta^{\text{raw}}(y_t \mid x, y_{<t})}{\mu^{\text{raw}}(y_t \mid x, y_{<t})}.$$

Consequently, whenever top- k /top- p truncation is active, $\hat{\rho}$ is not the exact importance ratio for either the unrestricted learner policy or the support-restricted policy $\pi_\theta^{\mathcal{S}}$. TETPO’s empirical ESS and the selected exponent α^* are therefore computed from $\hat{\rho}$ and should be viewed as using an approximate likelihood-ratio proxy.

This distinction affects the interpretation of our theoretical results. The ESS monotonicity, target-ESS feasibility, deterministic weight-dispersion bound, and positivity results (Theorems 2–3 and Proposition 2) hold for any strictly positive input ratios and therefore continue to hold exactly when TETPO is applied to $\hat{\rho}$. In contrast, the exact geometric-bridge identity and the associated bridge-target discrepancy results (Theorems 1 and 4 and Proposition 1) apply to the exact support-restricted ratio $\rho^{\mathcal{S}}$ and do not constitute guarantees for the approximate ratios used in the present experiments. Logging the retained support (or otherwise evaluating the post-truncation normalisers) would permit exact computation of $\rho^{\mathcal{S}}$; we leave this implementation and a direct comparison between exact and approximate ratios to future work.

Evaluation and overhead. Reported numbers are avg@16 accuracy (%) on the five benchmarks above: for each prompt we draw 16 independent samples with stochastic decoding (temperature 0.6, top- p 0.95, top- k 20, matching the rollout settings) and average the binary correctness over those samples. We use avg@16 rather than single-sample greedy accuracy because several of these benchmarks are small (AIME24/AIME25 contain 30 problems each), so a single greedy decode per

prompt gives a very coarse, high-variance estimate. Each configuration is run with three training seeds, and Table 1 reports mean $\pm$ standard deviation over those seeds; per-seed results accompany the code. The only computation TETPO adds to a GRPO step is the per-group 24-iteration bisection on already-computed log-ratios ($O(n_{\text{iter}} MG)$ elementwise operations) and the groupwise self-normalisation; these introduce no additional stored activations, and we observed no measurable change in per-step wall-clock relative to GRPO.

Appendix C. Additional Related Work

RL post-training and LLM agents. TETPO targets the off-policy correction that arises in asynchronous RL post-training of LLMs [2, 7, 8, 10], where group-relative objectives such as GRPO [9] and its sequence-level variants [12] are trained on stale rollouts. Beyond single-turn reasoning, RL is increasingly used to train multi-step LLM agents, where unstable updates are a central difficulty: hierarchical RL with stable temporal abstraction [3] stabilizes agent training by structuring decisions over temporally extended actions, and offline imagination with uncertainty-guided planning [4] reduces reliance on costly online interaction by planning over imagined trajectories. Both settings reuse data generated under a policy other than the current learner, so the weight-degeneracy problem studied here also arises there, and TETPO’s group-wise ESS control is complementary to these designs. A related line of work studies preference-based post-training and distillation, e.g. how the choice of models used to generate rejected responses affects preference distillation at scale [1]; such pipelines likewise train on samples drawn from a distribution that differs from the learner’s, which motivates distribution-mismatch corrections of the kind TETPO provides.