MHPO: MODULATED HAZARD-AWARE POLICY OPTIMIZATION FOR STABLE REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Controlling importance ratios in reinforcement learning with verifiable rewards requires retaining useful learning signals while limiting the influence of extreme policy deviations. Hard clipping can eliminate direct gradients for informative tokens, while unbounded ratio weighting can overemphasize outliers. We propose **M**odulated **H**azard-aware **P**olicy **O**ptimization (MHPO), which separates these concerns through its gradient multipliers. A Log-Fidelity Modulator (LFM) preserves local ratio fidelity and supplies a smooth, positive, globally bounded multiplier. A Decoupled Hazard Penalty (DHP) uses established cumulative hazard functions to attenuate ratio expansions and contractions separately, without reversing their direct gradient contributions. Under an explicit assumption on the joint moment of advantages and score norms, we derive a local bound on the second moment of the gradient. We also explain how repeated ratio contractions accumulate along autoregressive prefixes. Component ablations, token masking, and diagnostics of directional tails connect these design choices to retained learning signals and reduced prefix contraction. Among the evaluated baselines, MHPO achieves the best average across benchmarks on both Qwen3-4B-Base and Qwen3-30B-A3B-Base. Further evaluations on instruction-tuned and multimodal models support transfer across the tested settings.

1 INTRODUCTION

Reinforcement learning (RL) has become a central paradigm for the post-training of foundation models and unlocks extended chain-of-thought reasoning on complex mathematical and logical problems (Ouyang et al., 2022; OpenAI, 2024; Guo et al., 2025). Methods based on Group Relative Policy Optimization (GRPO) (Shao et al., 2024; Guo et al., 2025) have been scaled from small dense backbones to large Mixture-of-Experts (MoE) models, and are increasingly used as the default recipe for closing the gap between base models and the strongest reasoning-tuned counterparts. A parallel line of work has also extended GRPO beyond text, to vision-language models (Shen et al., 2025; Zhan et al., 2025; Yang et al., 2025b), which we cover in the appendix.

Despite these advances, ratio control presents a basic trade-off. In PPO/GRPO (Schulman et al., 2017; Shao et al., 2024), clipping removes the direct surrogate gradient when a token has already moved sufficiently far in its advantage-favored direction. This limits aggressive updates, but also discards potentially useful tail signals. Removing clipping restores those signals while exposing the update to large ratio-induced weights. DAPO (Yu et al., 2025), SAPO (Gao et al., 2025), CISPO (Chen et al., 2025), and MASPO (Fu et al., 2026) address this trade-off through different clipping or reweighting designs. Our question is how to retain a nonzero, advantage-aligned direct contribution while bounding its ratio-induced amplification and independently controlling directional attenuation.

Ratio direction adds a distinct consideration. We call $r_t > 1$ a *ratio expansion* and $r_t < 1$ a *ratio contraction*, both measured relative to the fixed behavior policy. These terms describe an existing probability displacement, not the advantage sign or the direction of the next update. Expanding a narrow set of tokens can concentrate learning. Contracting a useful token reduces the probability of every continuation sharing that prefix, holding the remaining conditional probabilities fixed. For autoregressive policies, the prefix likelihood ratio is $\exp(\sum_{j<t} \log r_j)$, so repeated contractions can

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.

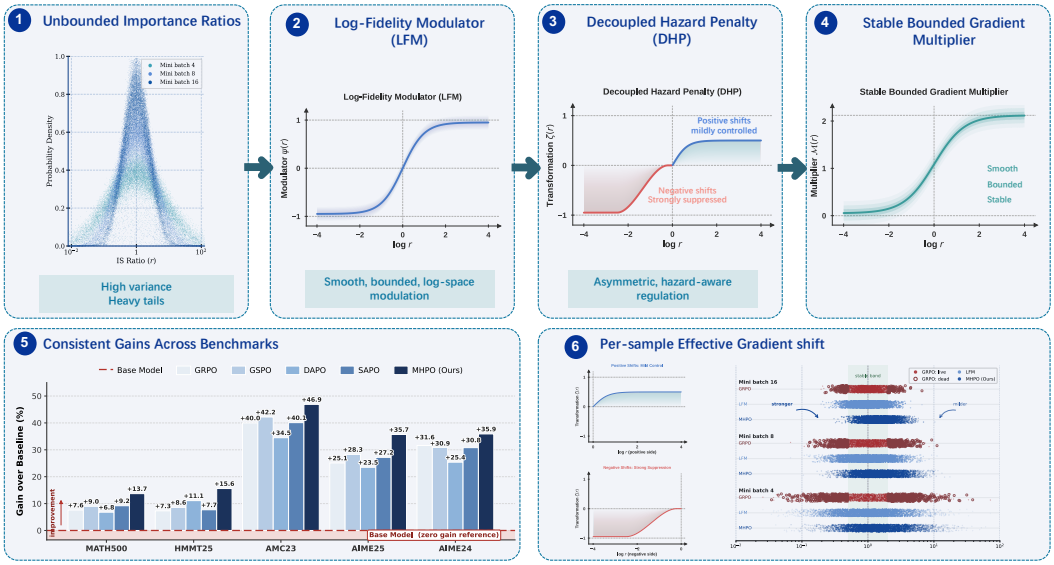


Figure 1: **Overview of MHPO. (1–4)** Ratio control proceeds from token importance ratios through log-space modulation and directional hazard attenuation to bounded gradient weighting. **(5)** Gains over Qwen3-4B-Base (Avg@32, percentage points) across five benchmarks. **(6)** Directional regulation and per-sample effective gradient shifts. Positive and negative shifts denote ratio expansion and contraction relative to the behavior policy. Sections 3 and 4.3 give the operator definitions and mechanism diagnostics.

compound even when each token-level change is modest. This multiplicative structure motivates separate control of expansion and contraction.

We propose **M**odulated **H**azard-aware **P**olicy **O**ptimization (MHPO) to separate signal retention from directional damping (Fig. 1). The Log-Fidelity Modulator (LFM) maps log ratios through a scaled $\tanh$, preserving the unclipped multiplier near the behavior-policy anchor while smoothly limiting outlier influence. The Decoupled Hazard Penalty (DHP) applies a detached survival weight from the established Weibull cumulative-hazard family (Cox & Oakes, 1984; Kalbfleisch & Prentice, 2002). Its positive weight attenuates each LFM contribution while preserving its direction. Separate branches allow stronger damping on the contraction side. Our contribution combines these operators and analyzes their gradient and prefix effects in policy optimization.

Our contributions are threefold. First, LFM and DHP provide positive bounded influence with separately configurable directional attenuation. Second, we derive a global-in-ratio multiplier bound and a local gradient moment certificate under a product-moment condition, together with the prefix identity that motivates contraction control. Third, evaluations and mechanism diagnostics connect the design to retained clipped-tail signals, reduced tail severity, and improved checkpoint retention. Appendix C gives the diagnostic definitions and derivations.

2 Related Work

Ratio-control methods for GRPO. PPO (Schulman et al., 2017) and GRPO (Shao et al., 2024) motivate many recent RLVR objectives. DAPO (Yu et al., 2025) changes clipping boundaries, SAPO (Gao et al., 2025) uses soft gating, and GSPO (Zheng et al., 2025) uses sequence-level ratios. CISPO (Chen et al., 2025) clips detached importance weights rather than dropping the direct learning signal outside a clip boundary. GMPO (Zhao et al., 2025) changes arithmetic aggregation to a geometric mean. MASPO (Fu et al., 2026) combines Gaussian soft gating, a probability-mass-adaptive limiter, and an asymmetric risk controller. MHPO shares the goal of retaining useful gradients with these methods. Its specific design combines a locally faithful log-ratio map with detached directional attenuation and a closed-form bound on their composite score multiplier.

Directional asymmetry. A complementary line distinguishes advantage polarities. SAPO (Gao et al., 2025) selects its sigmoid-gate temperature by advantage sign. MHPO instead separates a log-ratio modulator with a closed-form multiplier bound from detached attenuation conditioned on ratio direction. ASPO (Wang et al., 2026) modifies importance weighting asymmetrically across positive- and negative-advantage samples. Advantage-level approaches (Tang et al., 2025; Nan et al., 2025) reshape the learning signal itself. Ratio direction and advantage polarity are distinct: a token with negative advantage may still have $r > 1$ and require a corrective probability decrease.

Trust-region methods. TRPO's KL constraint (Schulman et al., 2015) and PPO's clipped surrogate (Schulman et al., 2017) are established approaches to controlling local policy changes. DPPO (Qi et al., 2026) uses distributional divergence to assess proximity, while LPO (Qu et al., 2026) formulates group-based RLVR as target projection on a response simplex. MHPO acts directly on sampled-token gradient multipliers, providing a complementary form of local update control. We compare against CISPO and MASPO in the same experimental framework.

3 METHODOLOGY

We first derive the GRPO gradient to identify the signal-retention problem, then introduce LFM and DHP, and finally connect their local update properties to autoregressive prefixes and the moment bound.

3.1 PRELIMINARIES OF GRPO

Let π_θ denote an autoregressive language model parameterized by θ. Given a query p from the query set $\mathcal{D}$, GRPO (Shao et al., 2024) samples a group of K responses $\{q^1, \ldots, q^K\}$ from the reference policy $\pi_{\theta_{\text{old}}}$. The likelihood of each response q^i is factorized as $\pi_{\theta_{\text{old}}}(q^i \mid p) = \prod_{t=1}^{T_i} \pi_{\theta_{\text{old}}}(q_t^i \mid p, q_{<t}^i)$, where T_i denotes the sequence length. Each response q^i is then evaluated by a reward model or a verifiable rule to obtain a scalar reward R^i.

Following (Shao et al., 2024), the GRPO training objective is defined at the token level as:

$$\mathcal{L}_{\text{GRPO}}(\theta) = -\mathbb{E}_{\substack{p\sim\mathcal{D}, \\ \{q^i\}_{i=1}^{K}\sim\pi_{\theta_{\text{old}}}(\cdot|p)}} \left[\frac{1}{K}\sum_{i=1}^{K} \frac{1}{T_i}\sum_{t=1}^{T_i} \min\Big(r_t^i(\theta)\,\hat{A}_t^i,\ \text{clip}\big(r_t^i(\theta),\,1-\epsilon,\,1+\epsilon\big)\,\hat{A}_t^i \Big) \right], \tag{1}$$

where

$$r_t^i(\theta) = \frac{\pi_\theta(q_t^i \mid p, q_{<t}^i)}{\pi_{\theta_{\text{old}}}(q_t^i \mid p, q_{<t}^i)}, \quad \hat{A}_t^i = \hat{A}^i = \frac{R^{(i)} - mean(\{R^1, \ldots, R^K\})}{std(\{R^1, \ldots, R^K\})}, \quad \epsilon > 0. \tag{2}$$

Here $\pi_{\theta_{\text{old}}}$ is the fixed behavior policy, not an optional KL-reference model. The group-standardized advantage $\hat{A}^i$ is shared across tokens of response i, not across different responses. We isolate the policy-surrogate term and omit any separate KL regularizer.

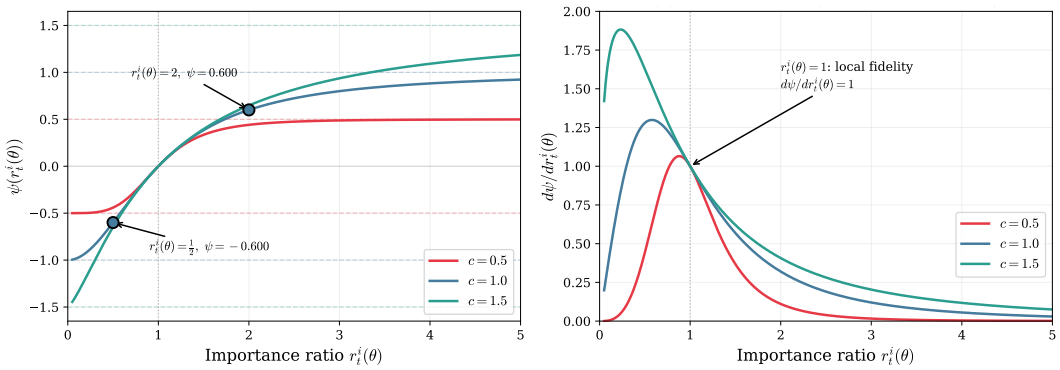
From the clipped objective to its gradient. Let $z_t = \log r_t$ and $s_t = \nabla_\theta \log \pi_\theta(q_t \mid p, q_{<t})$. Since the behavior policy is fixed, $\nabla_\theta r_t = r_t s_t$. Away from the clip boundaries, differentiating the per-token maximization objective gives

$$\begin{aligned} \nabla_\theta J_{\text{GRPO},t} &= \hat{A}_t \mathcal{M}_{\text{GRPO},t} s_t, \\ \mathcal{M}_{\text{GRPO},t} &= r_t \mathbf{1}[(\hat{A}_t > 0 \wedge r_t < 1+\epsilon) \vee (\hat{A}_t < 0 \wedge r_t > 1-\epsilon)]. \end{aligned} \tag{3}$$

Thus the *GRPO-dead set* is $D_t = \mathbf{1}[(\hat{A}_t > 0, r_t > 1+\epsilon) \vee (\hat{A}_t < 0, r_t < 1-\epsilon)]$. "Dead" means zero *direct surrogate contribution*, not a frozen token probability under shared parameters. For example, $r_t > 1$ with $\hat{A}_t < 0$ remains active and encourages a corrective decrease. Ratio direction and advantage sign must therefore be distinguished.

3.2 THE MHPO OBJECTIVE

We propose the MHPO objective, which is designed to enhance training stability and provide fine-grained control over gradient updates. The model parameters θ are optimized by minimizing the



(a) LFM Operator $\psi(r_t^i(\theta))$ (b) LFM Derivative (Gradient Modulation)

Figure 2: **Characteristics of the Log-Fidelity Modulator (LFM). (a)** The LFM operator $\psi(r_t^i(\theta))$ with different values of c. The transformation exhibits reciprocal antisymmetry ($\psi(1/r_t^i(\theta)) = -\psi(r_t^i(\theta))$) around the on-policy anchor ($r_t^i(\theta) = 1$) and smoothly bounds the output within the interval $[-c, c]$. The marked points at $r_t^i(\theta) = \frac{1}{2}$ and $r_t^i(\theta) = 2$ (for $c = 1.0$) confirm $\psi(\frac{1}{2}) = -\psi(2)$. **(b)** The derivative of the LFM map, with $d\psi/dr = 1$ at the anchor. The parameter-gradient multiplier includes the additional factor $re^{\psi(r)}$ in Eq. (7).

following objective function:

$$\mathcal{L}_{\text{MHPO}}(\theta) = -\mathbb{E}_{p\sim\mathcal{D},\ \{q^i\}_{i=1}^{K}\sim\pi_{\theta_{\text{old}}}(\cdot|p)}\left[\frac{1}{K}\sum_{i=1}^{K}\frac{1}{T_i}\sum_{t=1}^{T_i}\exp\Big(\psi(r_t^i(\theta)) - \text{sg}\big(\zeta(r_t^i(\theta))\big)\Big)\hat{A}_t^i\right], \quad (4)$$

where the transformation functions $\psi(\cdot)$ and $\zeta(\cdot)$ are defined as:

$$\psi(\cdot) = c\tanh\frac{\log(\cdot)}{c}, \quad \zeta(\cdot) = \left(\frac{s(\psi(\cdot))}{\lambda_+}\right)^{k_+} + \left(\frac{s(-\psi(\cdot))}{\lambda_-}\right)^{k_-}. \quad (5)$$

Here $\text{sg}(\cdot)$ is stop-gradient, $s(x) = \log(1 + e^x)$ is standard softplus (fixed smoothing $\beta = 1$), and $c, k_+, k_-, \lambda_+, \lambda_- > 0$ are the five configurable scalars. The implemented objective differentiates through LFM while detaching DHP. LFM acts on the ratio, not the reward. Its domain is $r > 0$, as softmax probabilities are positive for finite logits. We compute $z = \log\pi_\theta - \log\pi_{\theta_{\text{old}}}$ directly to avoid forming an underflowed ratio. We next elaborate on the two principal components of this objective, namely the Log-Fidelity Modulator and the Decoupled Hazard Penalty.

3.2.1 The Log-Fidelity Modulator

The Log-Fidelity Modulator (LFM) is defined as $\psi(\cdot) = c\tanh(\log(\cdot)/c)$. It projects the importance ratio $r_t^i(\theta)$ into log-space and bounds its contribution to the objective within $[-c, c]$. In contrast to the hard clipping used in standard GRPO, which produces gradient discontinuities and vanishing regions at the trust-region boundary, the LFM applies a smooth saturation that preserves differentiability across the entire domain, as shown in Figure 2(a). The derivative of the LFM operator is

$$\frac{d\psi}{dr_t^i(\theta)} = \frac{1}{r_t^i(\theta)}\text{sech}^2\left(\frac{\log r_t^i(\theta)}{c}\right), \quad (6)$$

The relevant quantity for the parameter gradient also includes $\nabla_\theta r = r s_t$. Writing $u(z) = c\tanh(z/c)$, its multiplier is

$$\nabla_\theta J_{\text{LFM},t} = \hat{A}_t \mathcal{M}_{\text{LFM}}(z_t) s_t, \qquad \mathcal{M}_{\text{LFM}}(z) = e^{u(z)}u'(z) = e^{c\tanh(z/c)}\text{sech}^2(z/c) > 0. \quad (7)$$

The $1/r$ factor cancels in this chain rule. Tail control concerns $\mathcal{M}_{\text{LFM}}$, not $d\psi/dr$ in isolation. The map $u(z)$ is strictly increasing, with $u''(z) = -(2/c)\text{sech}^2(z/c)\tanh(z/c)$. Its curvature changes sign at zero, so it is neither globally convex nor globally concave.

Local fidelity. As $z \to 0$, $u(z) = z + O(z^3)$ and $\mathcal{M}_{\text{LFM}}(z) = e^z(1 + O(z^2))$, recovering the unclipped ratio multiplier locally. **Bounded tail influence.** The multiplier is positive at every finite z, tends to zero in both tails, and has the finite supremum in Theorem 1. Unlike clipping, it has no finite dead interval. The token-masking results in Section 4.3 show the utility of the retained clipped-tail contributions. **Smoothness.** LFM is C^∞ on $r > 0$, removing clip-boundary discontinuities.

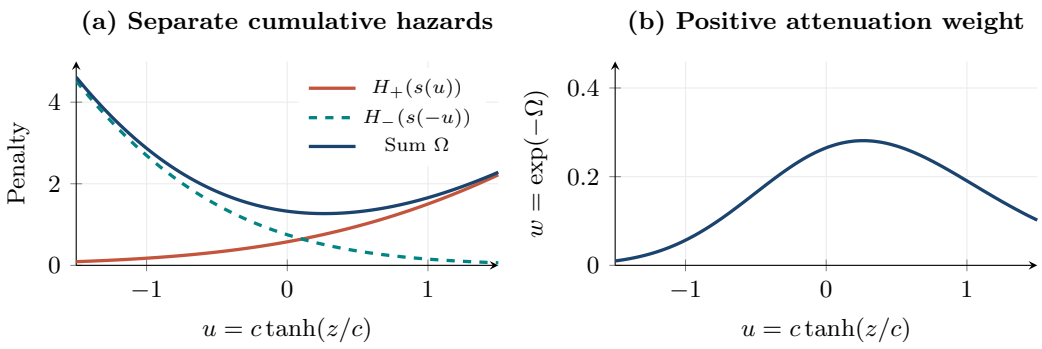


Figure 3: **DHP with the default asymmetric parameters. (a)** The separate cumulative hazards and their sum, with $(k_+, \lambda_+, k_-, \lambda_-) = (1.5, 1.0, 2.0, 0.8)$. **(b)** The corresponding weight $w = e^{-\Omega} \in (0, 1]$. Both panels use $c = 1.5$ and fixed $\beta = 1$. DHP attenuates more strongly on the contraction side for these settings.

3.2.2 The Decoupled Hazard Penalty

LFM bounds ratio-induced amplification. DHP adds directional control through a positive weight that preserves each direct contribution's direction and regulates expansion and contraction with separate hazard branches (Fig. 3):

$$\zeta(\cdot) = \left(\frac{s(\psi(\cdot))}{\lambda_+}\right)^{k_+} + \left(\frac{s(-\psi(\cdot))}{\lambda_-}\right)^{k_-}, \tag{8}$$

where the subscripts $+$ and $-$ refer to the signs of the modulated log ratio, not of $\hat{A}_t$. The softplus branches overlap near zero. In particular, $\zeta(1) > 0$, so DHP introduces baseline attenuation at the anchor rather than leaving it exactly unchanged.

Why cumulative hazards? In survival analysis, a survival function $S(x)$ and its cumulative hazard satisfy $H(x) = -\log S(x)$ (Cox & Oakes, 1984; Kalbfleisch & Prentice, 2002). The established Weibull family gives

$$H(x) = \left(\frac{x}{\lambda}\right)^k. \tag{9}$$

Thus $e^{-H(x)} \in (0, 1]$ is naturally an attenuation factor, λ sets the order-one scale, and $k > 1$ makes $H(x)/x$ increase with x. We use this family as a deterministic shaping operator. Since $u(z)$ is bounded, the composed magnitudes are also bounded. Increasing k strengthens attenuation where $x/\lambda > 1$ and weakens it where $x/\lambda < 1$, controlling the contrast between central and tail magnitudes.

How attenuation relates to prefix contraction. Write $\Omega(z) = \zeta(e^z)$ and $w(z) = e^{-\Omega(z)}$. For an isolated token, one SGD ascent step from the same θ, with the same advantage, score vector, and step size η, satisfies

$$\delta\theta_{\mathrm{LFM}} = \eta \hat{A}_t \mathcal{M}_{\mathrm{LFM}}(z_t) s_t, \qquad \delta\theta_{\mathrm{MHPO}} = w(z_t)\delta\theta_{\mathrm{LFM}}. \tag{10}$$

Consequently, for $\Delta z_X = z_t(\theta + \delta\theta_X) - z_t(\theta)$, local smoothness gives $\Delta z_{\mathrm{MHPO}} = w(z_t)\Delta z_{\mathrm{LFM}} + O(\eta^2)$. When the direct update further contracts the token, DHP reduces that contraction to first order. For a fixed prefix,

$$C_t = \log \frac{\pi_\theta(q_{\le t} \mid p)}{\pi_{\theta_{\mathrm{old}}}(q_{\le t} \mid p)} = \sum_{j \le t} z_j, \qquad R_t = e^{C_t}. \tag{11}$$

Repeated negative terms therefore multiply their effects on prefix likelihood. This motivates stronger contraction-side damping: reducing direct contributions that reinforce contraction addresses a token-level source of cumulative prefix suppression.

With shared parameters, one mini-batch SGD step gives the evaluated token j

$$\Delta z_j = \eta \sum_t \alpha_t \hat{A}_t \mathcal{M}_{\mathrm{LFM}}(z_t) w(z_t)\, s_j^\top s_t + O(\eta^2), \tag{12}$$

Algorithm 1 MHPO Training

Require: Policy π_θ, dataset $\mathcal{D}$, LFM bound c, DHP parameters $(k_+, \lambda_+, k_-, \lambda_-)$

for each iteration **do**
  Freeze $\theta_{\text{old}} \leftarrow \theta$; sample prompts and K responses per prompt using $\pi_{\theta_{\text{old}}}$
  Compute group-normalized advantages $\hat{A}_t^i$
  for each actor mini-batch update on this rollout **do**
    for each response q^i, each token t **do**
      $z_t^i \leftarrow \log \pi_\theta(q_t^i \mid p, q_{<t}^i) - \log \pi_{\theta_{\text{old}}}(q_t^i \mid p, q_{<t}^i)$
      $\psi_t \leftarrow c \tanh(z_t^i / c)$ *// LFM, computed directly in log space*
      $\zeta_t \leftarrow \big(s(\text{sg}(\psi_t))/\lambda_+\big)^{k_+} + \big(s(-\text{sg}(\psi_t))/\lambda_-\big)^{k_-}$ *// DHP*
    end for
    Average $-\exp(\psi_t - \zeta_t)\hat{A}_t^i$ with the weights in Eq. (4); update θ
  end for
end for

where α_t are normalized batch weights. The inner products $s_j^\top s_t$ couple token updates, and DHP weights each contribution separately. This connects directional attenuation to realized probability changes through shared representations. Section 4.3 measures the resulting tail and prefix effects, including a uniform attenuation control. Appendix C.2 gives the derivation.

3.3 Stability Analysis of MHPO

The stop-gradient in Eq. (4) is essential to the direction-preserving design. Full differentiation would instead yield $e^{u-\Omega}(u' - \Omega')$, whose sign can reverse for some parameter settings. Detaching the hazard gives a positive multiplier with bounded ratio-induced amplification.

The token-level MHPO objective is

$$J_{\text{MHPO}}(r_t^i(\theta)) = \exp\big(\psi(r_t^i(\theta)) - \text{sg}\big(\zeta(r_t^i(\theta))\big)\big)\,\hat{A}_t^i, \tag{13}$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operator. Applying the chain rule (Appendix B.3) yields

$$\nabla_\theta J_{\text{MHPO}}(r_t^i(\theta)) = \underbrace{\exp\big(\psi(r_t^i(\theta)) - \zeta(r_t^i(\theta))\big)\,\text{sech}^2\Big(\tfrac{\log r_t^i(\theta)}{c}\Big)}_{\mathcal{M}(r_t^i(\theta))}\,\nabla_\theta \log \pi_\theta(q_t^i \mid p, q_{<t}^i)\,\hat{A}_t^i, \tag{14}$$

where $\mathcal{M}(r)$ is the gradient multiplier that smoothly rescales the per-token gradient. For compactness, $\zeta(\cdot)$ denotes $\text{sg}(\zeta(\cdot))$ in the equation above.

The certificate propagates a pointwise multiplier bound to a normalized mini-batch gradient under an explicit local moment assumption. It requires neither token independence nor a factorization of advantage and score moments.

Theorem 1 (MHPO multiplier-to-moment stability certificate)**.** *Consider the normalized mini-batch semi-gradient* $g(\theta) = K^{-1}\sum_{i=1}^{K} T_i^{-1}\sum_{t=1}^{T_i} \mathcal{M}(r_t^i(\theta))\,\nabla_\theta \log \pi_\theta(q_t^i \mid p, q_{<t}^i)\,\hat{A}_t^i$*. At the current parameter value, write* $s_t^i = \nabla_\theta \log \pi_\theta(q_t^i \mid p, q_{<t}^i)$ *and assume* $\mathbb{E}_\mu[K^{-1}\sum_i T_i^{-1}\sum_t (\hat{A}_t^i)^2 \|s_t^i\|^2] \le C$*. Define the exact LFM multiplier certificate* $M_\psi(c) = \frac{2}{1+\sqrt{1+c^2}}\exp\Big(\frac{c^2}{1+\sqrt{1+c^2}}\Big)$*. Then, for all* $r > 0$ *and* $\tau > 0$*,*

$$0 \le \mathcal{M}(r) \le M_\psi(c),$$
$$\mathbb{E}_\mu\big[\|g(\theta)\|^2\big] \le C M_\psi(c)^2, \quad \Pr_\mu(\|g(\theta)\| \ge \tau) \le \frac{C M_\psi(c)^2}{\tau^2}. \tag{15}$$

Proof sketch. Since $0 < w \le 1$, $\mathcal{M} \le e^u u' \le M_\psi(c)$. Convexity gives $\|g\|^2 \le K^{-1}\sum_i T_i^{-1}\sum_t \mathcal{M}_t^2 (\hat{A}_t^i)^2 \|s_t^i\|^2$. Taking expectations and applying Markov's inequality proves the result. The multiplier bound is global in ratio, and the gradient moment bound applies at the current parameter value under the stated condition on C. DHP shapes directional influence while preserving the c-only multiplier bound. Full proofs appear in Appendix B.

Table 1: **Main results on the Qwen3-4B-Base foundation model without instruction tuning.** We report Avg@32 (%) for the best checkpoint. Avg. denotes the macro-average across benchmarks.

Method	MATH500	HMMT25	AMC23	AIME25	AIME24	Avg.
Baseline	57.2 (+0.0)	3.8 (+0.0)	41.3 (+0.0)	5.7 (+0.0)	8.8 (+0.0)	23.4 (+0.0)
GRPO	64.8 (+7.6)	11.1 (+7.3)	81.3 (+40.0)	30.8 (+25.1)	40.4 (+31.6)	45.7 (+22.3)
GSPO	66.2 (+9.0)	12.4 (+8.6)	83.5 (+42.2)	34.0 (+28.3)	39.7 (+30.9)	47.2 (+23.8)
DAPO	64.0 (+6.8)	14.9 (+11.1)	75.8 (+34.5)	29.2 (+23.5)	34.2 (+25.4)	43.6 (+20.2)
SAPO	66.4 (+9.2)	11.5 (+7.7)	81.4 (+40.1)	32.9 (+27.2)	39.6 (+30.8)	46.4 (+23.0)
CISPO	69.0 (+11.8)	11.0 (+7.2)	85.3 (+44.0)	33.5 (+27.8)	43.7 (+34.9)	48.5 (+25.1)
MASPO	67.5 (+10.3)	12.8 (+9.0)	82.6 (+41.3)	34.5 (+28.8)	40.8 (+32.0)	47.6 (+24.2)
MHPO (Ours)	**70.9** (+13.7)	**19.4** (+15.6)	**88.2** (+46.9)	**41.4** (+35.7)	**44.7** (+35.9)	**52.9** (+29.5)

4 EXPERIMENTS

We evaluate Qwen3-4B-Base and Qwen3-30B-A3B-Base, with ablations and mechanism diagnostics on Qwen3-4B. Section 4.4 evaluates a smaller distilled reasoning model. Table 5(A–C) reports general-purpose, math-specialized, and vision-language instruction-tuned backbones.

4.1 EXPERIMENTAL SETUP

Models and datasets. The main results are reported on two foundation models of different scales and architectures, namely the dense Qwen3-4B-Base (Yang et al., 2025a) and the Mixture-of-Experts Qwen3-30B-A3B-Base (Yang et al., 2025a), both without prior instruction tuning. The policy is optimized on DAPO-Math-17k-Processed (Yu et al., 2025) with verifiable outcome-based rewards, and zero-shot generalization is evaluated on AIME 2024/2025 (Mathematical Association of America, 2025a), AMC 2023 (Mathematical Association of America, 2025b), HMMT25 (HMMT Organization, 2025), and MATH-500 (Hendrycks et al., 2021). The ablation studies in Section 4.3 are conducted on Qwen3-4B-Base under the same training and evaluation protocol. Additional instruction-tuned and vision-language results appear in Appendix D.

Baselines and metrics. We compare against GRPO, GSPO, DAPO, SAPO, CISPO, and MASPO. Avg@32 is the mean correctness of 32 sampled responses per prompt, averaged across prompts. It estimates single-sample pass@1, not the probability of at least one success in 32 attempts. Repeated-sample averaging is also used in DeepSeek-R1 and DAPO (Guo et al., 2025; Yu et al., 2025). Under independent samples with correctness probability p, the conditional sampling variance is $p(1-p)/32$. Appendix A gives the formula. "Avg." in tables is the unweighted mean across the displayed tasks.

Configuration and cost. We use $(k_+, \lambda_+, k_-, \lambda_-) = (1.5, 1.0, 2.0, 0.8)$, with $c = 1.5$ for the evaluated dense models and $c = 4.0$ for Qwen3-MoE. On the H20 training platform, the additional work consists of $O(BT)$ token-wise scalar operations for B responses of length T, with no extra rollout, model forward pass, or trainable module. Appendix A gives implementation details.

4.2 MAIN RESULTS

Tables 1 and 2 show the best macro-average for MHPO at both scales. On Qwen3-4B, MHPO reaches 52.9%, exceeding CISPO by 4.4 points and MASPO by 5.3 points. It leads on all five tasks. Among baselines, DAPO is best on HMMT25 (14.9%), while MASPO is best on AIME25 (34.5%). MHPO's gains over these task-specific baselines are 4.5 and 6.9 points. On Qwen3-MoE, MHPO reaches 54.7% versus CISPO's 54.0% and MASPO's 49.7%. CISPO remains higher on HMMT25 (37.8% versus 37.3%), and MHPO leads on the other four tasks.

4.3 ABLATION STUDIES

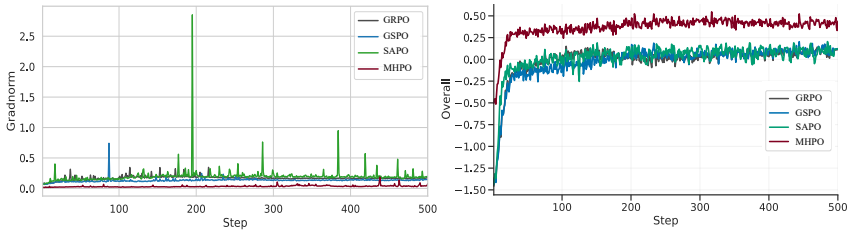
We study component utility, parameter sensitivity, and mechanism diagnostics on Qwen3-4B-Base. Component ablations establish which parts help performance, while the diagnostics test the specific signal-retention and directional-contraction explanations.

Table 2: **Main results on the Qwen3-30B-A3B-Base foundation model without instruction tuning.** We report Avg@32 (%) for the best checkpoint. Avg. denotes the macro-average across benchmarks.

Method	MATH500	HMMT25	AMC23	AIME25	AIME24	Avg.
Baseline	76.5 (+0.0)	1.6 (+0.0)	43.6 (+0.0)	9.0 (+0.0)	14.2 (+0.0)	29.0 (+0.0)
GRPO	82.8 (+6.3)	35.3 (+33.7)	78.4 (+34.8)	28.9 (+19.9)	31.6 (+17.4)	51.4 (+22.4)
GSPO	84.1 (+7.6)	36.9 (+35.3)	80.6 (+37.0)	29.8 (+20.8)	32.4 (+18.2)	52.8 (+23.8)
DAPO	84.5 (+8.0)	37.2 (+35.6)	81.0 (+37.4)	30.2 (+21.2)	32.8 (+18.6)	53.1 (+24.2)
SAPO	83.2 (+6.7)	34.6 (+33.0)	76.8 (+33.2)	30.7 (+21.7)	29.4 (+15.2)	50.9 (+22.0)
CISPO	85.4 (+8.9)	**37.8** (+36.2)	82.3 (+38.7)	31.6 (+22.6)	33.1 (+18.9)	54.0 (+25.1)
MASPO	81.7 (+5.2)	33.8 (+32.2)	75.9 (+32.3)	28.1 (+19.1)	28.8 (+14.6)	49.7 (+20.7)
MHPO (Ours)	**86.6** (+10.1)	37.3 (+35.7)	**83.6** (+40.0)	**32.4** (+23.4)	**33.8** (+19.6)	**54.7** (+25.8)

Table 3: **Ablation analysis on Qwen3-4B-Base.** Top: sensitivity to c, k, and λ, and component/directional DHP variants (macro-average Avg@32, %). Bottom: gradient-norm and overall-reward trajectories during training.

(a) Bound c (hazard-free)		*(b) Shape k* (c=1.5, λ=0.8)		*(c) Scale λ* (c=1.5, k=3.0)		*(d) Components / DHP direction*	
c	Avg@32	k	Avg@32	λ	Avg@32	Variant	Avg@32
0.5	50.9	1.0	52.0	0.5	52.5	DHP only	50.8
1.0	51.3	1.5	52.3	**0.8**	**52.9**	LFM only	52.0
1.5	**52.0**	2.5	52.6	1.0	52.7	Pos.-stronger DHP	52.1
2.0	51.7	**3.0**	**52.9**	2.0	51.9	Symmetric DHP	52.4
						Neg.-stronger DHP	**52.9**



(a) Gradnorm (b) Overall reward

Hyperparameter sensitivity. Table 3 reports Qwen3-4B-Base ablations using macro-average Avg@32 across the five math benchmarks. The table sweeps the LFM bound c, Weibull shape k, and Weibull scale λ, and further isolates the two components and the directionality of the DHP. These parameters correspond to the two design axes of MHPO: c controls the fidelity-preserving ratio range, while k and λ control the sharpness and onset of tail damping. For the k and λ sweeps, other hyperparameters are fixed to the reference setting; for the c sweep, the hazard term is disabled to isolate the smooth bound.

LFM alone reaches 52.0% and DHP alone 50.8%, compared with 52.9% for their combination. The ratio-contraction-stronger variant also exceeds the symmetric (52.4%) and expansion-stronger (52.1%) alternatives. These differences support the complementary roles of the two components and stronger contraction-side control. The hazard-free c sweep spans 50.9–52.0%, and the reported k and λ sweeps span 52.0–52.9% and 51.9–52.9%. Performance remains within these ranges across the tested one-factor settings.

Training stability. The lower panel of Table 3 shows gradient norms and rewards. MHPO's best-to-latest Avg@32 changes from 52.9% to 52.1%, a 0.8-point drop, compared with 6.9–12.2 points for the evaluated baselines (Table 6). The following diagnostics relate this improved checkpoint retention to the token contributions and prefix dynamics targeted by LFM and DHP.

Mechanism evidence. The GRPO-dead set comprises 18.6% of nonzero-advantage tokens and 27.4% of absolute-advantage mass. LFM gives this set 34.7% of its effective coefficient mass, defined

as $\sum_t D_t|\hat{A}_t|\mathcal{M}_{\text{LFM}}(z_t)/\sum_t|\hat{A}_t|\mathcal{M}_{\text{LFM}}(z_t)$. This measures the retained scalar weight entering the policy gradient. The advantage-aligned log-probability change, $\text{sign}(\hat{A}_t)(\log p_{\text{after}} - \log p_{\text{before}})$, is 0.019 for LFM versus 0.002 for GRPO. Removing these specific contributions reduces LFM's best/latest Avg@32 from $52.0/48.5$ to $49.2/42.6$, while a shuffled mask matched for count, position, advantage sign, $|A|$ bin, and removed coefficient mass gives $51.1/46.8$. The larger loss from the exact mask shows that the recovered clipped-tail signals contribute to both task performance and checkpoint retention.

DHP reduces the fraction below the negative GRPO clip boundary from 8.9% with LFM to 3.9% with MHPO. A scalar attenuation control matched to MHPO's update norm remains at 8.0%. Negative-tail 1%-CVaR, the mean contraction severity $(-\log r)_+$ in the worst 1% of tokens, falls from 1.18 to 0.63, versus 1.09 for the scalar control. MHPO also retains more expansion-side variation than symmetric DHP: upper-boundary exceedance is 7.1% versus 5.8%, with highest-1% mean log ratio 0.91 versus 0.77. On correct responses, the fraction whose prefix likelihood ratio ever falls below 0.1 decreases from 13.2% with LFM and 11.5% with the control to 4.9% with MHPO. Together, the directional-tail contrast and reduced collapse of correct-response prefixes support selective contraction control beyond uniform shrinkage. Appendix C gives the definitions and local derivation.

Tail frequency measures how often contraction exceeds the clip boundary, while CVaR measures its severity in the worst tail. For example, CVaR 1.18 corresponds to a geometric mean token ratio of $e^{-1.18} \approx 0.307$ within that tail.

The first-crossing depth is $\tau_{0.1}(i) = \inf\{t : C_{i,t} < \log 0.1\}$, using the prefix log ratio in Eq. (11). Among correct responses that cross this threshold, the median depth is 846 tokens with LFM and $1{,}487$ with MHPO. Noncrossing responses have $\tau_{0.1} = \infty$ and are excluded from this conditional median. Reporting it alongside the collapse rate captures both the incidence of prefix suppression and its location within affected responses.

4.4 Transfer to a Distilled Reasoning Model

Table 4: **DeepSeek-R1-Distill-Qwen-1.5B results.** Avg@32 (%) and arithmetic mean across the four displayed tasks.

Method	AIME24	AIME25	AMC23	MATH500	4-task Avg.
GRPO	33.2	27.7	79.5	77.6	54.5
SAPO	39.3	28.0	**82.7**	79.1	57.3
MASPO	41.0	28.4	82.2	78.0	57.4
MHPO	**41.6**	**29.7**	82.5	**79.6**	**58.4**

Table 4 evaluates DeepSeek-R1-Distill-Qwen-1.5B under the same Avg@32 protocol. MHPO reaches 58.4%, exceeding MASPO by 1.0 point in the four-task average. It leads on AIME24, AIME25, and MATH500, while SAPO remains best on AMC23. On Qwen2.5-Instruct, Math, and VL backbones, MHPO scores $43.7/47.8/53.0$, versus DAPO's $40.2/43.6/52.3$ (Table 5). These evaluations cover Qwen-derived dense, MoE, instruction-tuned, distilled, and vision-language backbones, supporting transfer across model scales, training lineages, and modalities.

5 Conclusion

MHPO separates retained learning signals from directional attenuation through LFM and DHP. Its positive multiplier preserves direct update directions and bounds ratio-induced amplification, yielding a local gradient moment certificate under the stated product-moment condition. Token interventions show that recovered clipped-tail contributions support learning, while directional-tail and prefix diagnostics connect asymmetric damping to reduced contraction of correct-response prefixes. This combination achieves the best benchmark average at both evaluated Qwen3 scales and improves checkpoint retention, with further gains on instruction-tuned, distilled, and vision-language backbones.

AI USE STATEMENT

Generative AI tools assisted with refining the mechanistic explanations and mathematical presentation, checking derivations, designing diagnostic analyses, interpreting results, identifying relevant literature, revising manuscript text, preparing analytical figures, and LaTeX formatting. The empirical results reported here were confirmed by the authors from their experiments. The authors are responsible for the final text, mathematical claims, citations, and experimental evidence.

REPRODUCIBILITY STATEMENT

Section 3 specifies the objective, gradient multiplier, and update algorithm. Appendix A records the operator settings and training and evaluation protocol, including the definition of Avg@32. Appendix B states the theoretical assumptions and proofs. Appendix C defines the token, directional-tail, and prefix diagnostics and explains the scope of the local update analysis. Additional backbone and checkpoint results appear in Appendices D and E.

REFERENCES

Aili Chen, Aonian Li, Bangwei Gong, Binyang Jiang, Bo Fei, Bo Yang, Boji Shan, Changqing Yu, Chao Wang, Cheng Zhu, et al. Minimax-m1: Scaling test-time compute efficiently with lightning attention. *arXiv preprint arXiv:2506.13585*, 2025.

David R. Cox and David Oakes. *Analysis of Survival Data*. Chapman and Hall, 1984. doi: 10.1201/9781315137438.

Xiaoliang Fu, Jiaye Lin, Yangyi Fang, et al. Maspo: Unifying gradient utilization, probability mass, and signal reliability for robust and sample-efficient llm reasoning. *arXiv preprint arXiv:2602.17550*, 2026.

Chang Gao, Chujie Zheng, Xiong-Hui Chen, et al. Soft adaptive policy optimization. *arXiv preprint arXiv:2511.20347*, 2025.

Daya Guo, Dejian Yang, Haowei Zhang, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Dan Hendrycks, Collin Burns, Saurav Kadavath, et al. Measuring mathematical problem solving with the MATH dataset. In *Advances in Neural Information Processing Systems*, 2021.

HMMT Organization. Harvard-MIT mathematics tournament (HMMT). `https://www.hmmt.org/`, 2025. Annual mathematics tournament organized by Harvard and MIT students.

John D. Kalbfleisch and Ross L. Prentice. *The Statistical Analysis of Failure Time Data*. Wiley, 2 edition, 2002. doi: 10.1002/9781118032985.

Pan Lu, Ran Gong, Shibiao Jiang, et al. Inter-GPS: Interpretable geometry problem solving with formal language and symbolic reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, 2021.

Pan Lu, Hritik Bansal, Tony Xia, et al. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations*, 2024.

Mathematical Association of America. American invitational mathematics examination (AIME). `https://maa.org/maa-invitational-competitions/`, 2025a. Annual invitational competition for top-scoring AMC participants.

Mathematical Association of America. American mathematics competitions (AMC). `https://maa.org/amc/`, 2025b. Annual competition series for students in grades 8–12.

Gongrui Nan, Siye Chen, Jing Huang, et al. NGRPO: Negative-enhanced group relative policy optimization. *arXiv preprint arXiv:2509.18851*, 2025.

OpenAI. Learning to reason with llms, 2024. URL https://openai.com/index/learning-to-reason-with-llms/.

Long Ouyang, Jeffrey Wu, Xu Jiang, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 2022.

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

Yun Qu, Qi Wang, Yixiu Mao, Heming Zou, Yuhang Jiang, Yingyue Li, Wutong Xu, Lizhou Cai, Weijie Liu, Clive Bai, Kai Yang, Yangkun Chen, Saiyong Yang, and Xiangyang Ji. Listwise policy optimization: Group-based RLVR as target-projection on the LLM response simplex. *arXiv preprint arXiv:2605.06139*, 2026. URL https://arxiv.org/abs/2605.06139.

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

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Haozhan Shen, Peng Liu, Jingcheng Li, et al. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv: 2504.07615*, 2025.

Xinyu Tang, Yuliang Zhan, Zhixun Li, et al. Rethinking sample polarity in reinforcement learning with verifiable rewards. *arXiv preprint arXiv:2512.21625*, 2025.

Jiakang Wang, Runze Liu, Qingpeng Cai, Lei Lin, Wenping Hu, Xiu Li, Fuzheng Zhang, Guorui Zhou, Kun Gai, and Ling Pan. When importance sampling misallocates credit: Asymmetric ratios for outcome-supervised RL. *arXiv preprint arXiv:2510.06062*, 2026. URL https://arxiv.org/abs/2510.06062v2.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with MATH-Vision dataset. *arXiv preprint arXiv:2402.14804*, 2024.

An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Yi Yang, Xiaoxuan He, Hongkun Pan, et al. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv: 2503.10615*, 2025b.

Qiying Yu, Zheng Zhang, Ruofei Zhu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Yufei Zhan, Yousong Zhu, Shurong Zheng, et al. Vision-r1: Evolving human-free alignment in large vision-language models via vision-guided reinforcement learning. *arXiv preprint arXiv: 2503.18013*, 2025.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, et al. MathVerse: Does your multi-modal LLM truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, 2024.

Yuzhong Zhao, Yue Liu, Junpeng Liu, Jingye Chen, Xun Wu, Yaru Hao, Tengchao Lv, Shaohan Huang, Lei Cui, Qixiang Ye, Fang Wan, and Furu Wei. Geometric-mean policy optimization. *arXiv preprint arXiv:2507.20673*, 2025. URL https://arxiv.org/abs/2507.20673.

Chujie Zheng, Shixuan Liu, Mingze Li, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

A Implementation Details

This section reports the operator settings and training/evaluation setup used in the main paper and supplementary experiments.

MHPO hyperparameters. The five scalars have distinct roles: c controls the log-ratio fidelity range, $k_\pm$ the directional shape, and $\lambda_\pm$ the attenuation scale. For the five original backbones, the DHP settings are fixed at $(k_+, \lambda_+) = (1.5, 1.0)$ and $(k_-, \lambda_-) = (2.0, 0.8)$. We use $c = 1.5$ on Qwen3-4B and the three Qwen2.5 instruction-tuned variants, and $c = 4.0$ on Qwen3-30B-A3B. The larger MoE value permits a wider log-ratio fidelity range. Softplus smoothing is fixed at $\beta = 1$ throughout.

Training data and protocol. The policy is optimized on DAPO-Math-17k-Processed (Yu et al., 2025) with verifiable outcome-based rewards, following the standard GRPO pipeline. We sample $K = 16$ responses per prompt during rollout for policy updates, and the maximum response length is 4,096 tokens for the Qwen3 foundation models (both the 4B dense and the 30B-A3B MoE) and 2,048 tokens for the Qwen2.5 variants used in the supplementary experiments. Reward normalization, advantage computation, and the underlying actor update follow standard GRPO practice.

Evaluation protocol. For math benchmarks (MATH-500 (Hendrycks et al., 2021), HMMT25 (HMMT Organization, 2025), AMC 2023 (Mathematical Association of America, 2025b), AIME 2024/2025 (Mathematical Association of America, 2025a)), we sample $n = 32$ responses per prompt and report Avg@32. For the vision-language benchmarks in Section D (MathVista (Lu et al., 2024), MathVision (Wang et al., 2024), MathVerse (Zhang et al., 2024)), we use the same Avg@32 protocol. All evaluations are zero-shot with respect to the reported benchmarks; only DAPO-Math-17k-Processed is used for policy optimization on text backbones, and only Geometry3K (Lu et al., 2021) is used for the Qwen2.5-VL-7B-Instruct results in Subtable (C) of Table 5.

Baselines. The main baseline comparisons use the same codebase, training data, and total update budget. GRPO (Shao et al., 2024) uses $\epsilon = 0.2$, and the DAPO-style clip configuration (Yu et al., 2025) uses $(\epsilon_{\text{low}}, \epsilon_{\text{high}}) = (0.2, 0.28)$. SAPO (Gao et al., 2025) uses soft gating, GSPO (Zheng et al., 2025) sequence-level ratio control, and CISPO (Chen et al., 2025) detached clipped importance weights. MASPO (Fu et al., 2026) combines Gaussian gating, a mass-adaptive limiter, and asymmetric control. These comparisons evaluate the policy-update operators within a common training framework.

Meaning of Avg@32. For N evaluation prompts, let $X_{ij} \in \{0, 1\}$ denote correctness of completion j for prompt i. We report

$$\text{Avg@32} = \frac{100}{N} \sum_{i=1}^{N} \widehat{p}_i, \qquad \widehat{p}_i = \frac{1}{32} \sum_{j=1}^{32} X_{ij}. \tag{16}$$

If the completions are sampled independently with success probability p_i, then $\mathbb{E}[\widehat{p}_i \mid i] = p_i$ and $\text{Var}(\widehat{p}_i \mid i) = p_i(1 - p_i)/32$. Thus Avg@32 estimates the single-sample pass@1 of the specified sampling distribution with lower Monte Carlo variance than one draw per prompt. By comparison, pass@32 measures the probability of at least one successful completion, with per-prompt target $1 - (1 - p_i)^{32}$. Repeated-sample pass-rate estimation is used in DeepSeek-R1 and DAPO (Guo et al., 2025; Yu et al., 2025).

Computational cost. The training platform uses NVIDIA H20 GPUs. Relative to the existing actor loss, MHPO adds token-wise $\tanh$, softplus, powers, and exponentiation on already available log-probabilities. The work is $O(BT)$ and can require $O(BT)$ temporary scalar storage, with no extra model evaluation, rollout, or trainable state.

B Theoretical Analysis and Proofs

This appendix is organized to mirror the theoretical narrative in the main paper. We first introduce the generic log-form view of ratio operators, then define the hazard-shaping components used by

MHPO, next establish the gradient identities and boundedness certificates that support the stability claims, and finally explain why the semi-gradient implementation is adopted in place of the fully differentiable variant. For self-containedness, we restate the main multiplier-to-moment certificate (Theorem 2 below) before giving its proof.

Restatement of the main stability certificate. Let $\psi(\cdot) = c\tanh(\log(\cdot)/c)$ be the Log-Fidelity Modulator (LFM) and let $\zeta(\cdot)$ be the Decoupled Hazard Penalty (DHP) composed of Weibull cumulative hazards on one-sided softplus magnitudes, as used in the main paper. Under the semi-gradient implementation (a stop-gradient is applied to ψ inside ζ), the MHPO gradient multiplier is

$$\mathcal{M}(r) = \exp\big(\psi(r) - \zeta(r)\big)\,\mathrm{sech}^2\left(\frac{\log r}{c}\right). \tag{17}$$

Theorem 2 (MHPO multiplier-to-moment stability certificate; restated from the main paper)**.** *Consider the normalized mini-batch semi-gradient* $g(\theta) = K^{-1}\sum_{i=1}^{K} T_i^{-1}\sum_{t=1}^{T_i}\mathcal{M}(r_t^i(\theta))s_t^i\hat{A}_t^i$, *where* $s_t^i = \nabla_\theta \log \pi_\theta(q_t^i \mid p, q_{<t}^i)$. *At the current parameter value, assume the normalized product moment*

$$\mathbb{E}_\mu\left[\frac{1}{K}\sum_i \frac{1}{T_i}\sum_t (\hat{A}_t^i)^2\|s_t^i\|^2\right] \le C. \tag{18}$$

Define the exact LFM multiplier certificate

$$M_\psi(c) = \frac{2}{1+\sqrt{1+c^2}}\exp\left(\frac{c^2}{1+\sqrt{1+c^2}}\right).$$

Then, for all $r > 0$ *and* $\tau > 0$,

$$\begin{gathered} 0 \le \mathcal{M}(r) \le M_\psi(c), \\ \mathbb{E}_\mu\big[\|g(\theta)\|^2\big] \le C M_\psi(c)^2, \quad \Pr_\mu(\|g(\theta)\| \ge \tau) \le \frac{C M_\psi(c)^2}{\tau^2}. \end{gathered} \tag{19}$$

The proof is given in Section B.5 below, after the supporting lemmas.

B.1 Formal Equivalence of the Log-Form Surrogate

Proposition 3 (Any positive ratio operator admits a log-form)**.** *Let* $\omega : (0,\infty) \to (0,\infty)$ *be any positive ratio operator applied to the importance ratio* $r = \exp(z)$. *Define* $\phi : \mathbb{R} \to \mathbb{R}$ *by* $\phi(z) \triangleq \log \omega(\exp(z))$. *Then* $\omega(r) = \exp(\phi(\log r))$, *i.e., the per-token ratio operator can always be written as* $\exp(\phi(z))$ *in log-ratio space.*

Proof. Substitute $r = \exp(z)$ to obtain $\exp(\phi(z)) = \exp(\log \omega(\exp(z))) = \omega(\exp(z)) = \omega(r)$. □

B.2 Definitions and Properties of Weibull Hazard Shaping

For an absolutely continuous survival law, $S(x) = \Pr(T > x)$, $H(x) = -\log S(x)$, and $h(x) = H'(x)$ are respectively the survival function, cumulative hazard, and instantaneous hazard (Cox & Oakes, 1984; Kalbfleisch & Prentice, 2002). Weibull survival has $S(x) = \exp[-(x/\lambda)^k]$. MHPO uses this established functional family for deterministic attenuation of directional ratio magnitudes.

Let $u(z) \triangleq c\tanh(z/c)$ and $\beta > 0$ (fixed to 1 in the implementation). We define ratio-expansion/contraction magnitudes using softplus:

$$x_+(u) = \frac{\mathrm{softplus}(\beta u)}{\beta} = \frac{\log(1+e^{\beta u})}{\beta}, \tag{20}$$

$$x_-(u) = \frac{\mathrm{softplus}(-\beta u)}{\beta} = \frac{\log(1+e^{-\beta u})}{\beta}. \tag{21}$$

These satisfy $x_+(u), x_-(u) > 0$ and $x_+(u) + x_-(u) = |u| + 2\log(1+e^{-\beta|u|})/\beta$, approaching $|u|$ at large magnitudes. At the anchor, both branches equal $\log 2/\beta$, giving finite baseline attenuation. Because $|u| \le c$, both magnitudes lie in the compact interval $[\log(1+e^{-\beta c})/\beta, \log(1+e^{\beta c})/\beta]$.

For $x \geq 0$, define the cumulative hazards

$$H_+(x) \triangleq \left(\frac{x}{\lambda_+}\right)^{k_+}, \qquad H_-(x) \triangleq \left(\frac{x}{\lambda_-}\right)^{k_-}, \tag{22}$$

with shape parameters $k_\pm > 0$ and scales $\lambda_\pm > 0$. The DHP composes $u \mapsto x_\pm(u) \mapsto H_\pm(x_\pm)$, yielding a nonnegative hazard penalty $\Omega(z) = H_+\big(x_+(u(z))\big) + H_-\big(x_-(u(z))\big)$ and survival weight $w(z) = \exp(-\Omega(z)) \in (0, 1]$.

B.3 Differentiation Identities and Gradient Bounds

Lemma 4 (Gradient form and gradient multiplier)**.** *For $\ell_t(\theta) \triangleq -A_t \exp(\phi(z_t(\theta)))$, the gradient satisfies*

$$\nabla_\theta \ell_t(\theta) = -A_t \, \exp(\phi(z_t(\theta))) \, \phi'(z_t(\theta)) \, \nabla_\theta z_t(\theta). \tag{23}$$

Equivalently, defining the gradient multiplier $\mathcal{M}_\phi(z) \triangleq \exp(\phi(z)) \, \phi'(z)$, we have $\nabla_\theta \ell_t(\theta) = -A_t \, \mathcal{M}_\phi(z_t(\theta)) \, \nabla_\theta z_t(\theta)$.

Proof. Since $\ell_t(\theta) = -A_t \, g(z_t(\theta))$ with $g(z) = \exp(\phi(z))$, the chain rule gives

$$\nabla_\theta \ell_t(\theta) = -A_t \, g'(z_t(\theta)) \, \nabla_\theta z_t(\theta).$$

Moreover, $g'(z) = \exp(\phi(z)) \, \phi'(z)$ by differentiating $g(z) = \exp(\phi(z))$. Substituting $g'(z_t(\theta))$ yields the claimed expression. □

Lemma 5 (Gradient of the log-ratio)**.** *With $z_t(\theta) = \log \pi_\theta(y_t \mid x, y_{<t}) - \log \pi_{\theta_{old}}(y_t \mid x, y_{<t})$, we have*

$$\nabla_\theta z_t(\theta) = \nabla_\theta \log \pi_\theta(y_t \mid x, y_{<t}). \tag{24}$$

Proof. The behavior policy $\pi_{\theta_{\text{old}}}$ is fixed during the update, so its log-probability term is constant w.r.t. θ. □

Proposition 6 (Per-token gradient bound via M)**.** *Assume $\|\nabla_\theta \log \pi_\theta(y_t \mid x, y_{<t})\| \leq G$ for all tokens in the update. Then the per-token gradient magnitude satisfies*

$$\|\nabla_\theta \ell_t(\theta)\| \leq |A_t| \; M \, G, \tag{25}$$

Proof. By Lemma 4, $\nabla_\theta \ell_t(\theta) = -A_t \exp(\phi(z_t(\theta))) \, \phi'(z_t(\theta)) \, \nabla_\theta z_t(\theta) = -A_t \, \mathcal{M}_\phi(z_t(\theta)) \, \nabla_\theta z_t(\theta)$. Lemma 5 gives $\nabla_\theta z_t(\theta) = \nabla_\theta \log \pi_\theta(y_t \mid x, y_{<t})$. Taking norms and using $\|\nabla_\theta \log \pi_\theta(\cdot)\| \leq G$ yields

$$\|\nabla_\theta \ell_t(\theta)\| \leq |A_t| \, |\mathcal{M}_\phi(z_t(\theta))| \, G \leq |A_t| \left(\sup_z |\mathcal{M}_\phi(z)|\right) G = |A_t| \, M \, G,$$

where we define $M \triangleq \sup_z |\mathcal{M}_\phi(z)|$. □

B.4 Log-Fidelity Modulator Derivatives and Proof of Theorem 8

Lemma 7 (Derivatives and bounds for $u(z) = c \tanh(z/c)$)**.** *For $u(z) = c \tanh(z/c)$ (equivalently, $u(z) = \psi(e^z)$ with the LFM ψ),*

$$u'(z) = \operatorname{sech}^2(z/c) \in (0, 1], \qquad |u''(z)| \leq \frac{2}{c}. \tag{26}$$

Proof. Differentiate $\tanh$ and use $|\tanh(\cdot)| \leq 1$ and $\operatorname{sech}^2(\cdot) \leq 1$. □

Theorem 8 (Closed-form $M_\psi(c)$ for the Log-Fidelity Modulator)**.** *Let $u(z) = c \tanh(z/c)$ with parameter $c > 0$, and define $M_\psi(c) \triangleq \sup_{z \in \mathbb{R}} \exp(u(z)) \, u'(z)$. Then*

$$M_\psi(c) = \frac{2}{1 + \sqrt{1 + c^2}} \exp\left(\frac{c^2}{1 + \sqrt{1 + c^2}}\right). \tag{27}$$

Proof of Theorem 8. Let $t = z/c$. Then $u(z) = c\tanh t$ and $u'(z) = \operatorname{sech}^2 t$. We maximize $f(t) \triangleq \exp(c\tanh t)\operatorname{sech}^2 t$ over $t \in \mathbb{R}$. Differentiate $\log f(t) = c\tanh t + \log(\operatorname{sech}^2 t)$ to obtain $\frac{d}{dt}\log f(t) = c\operatorname{sech}^2 t - 2\tanh t$. Setting it to zero gives $c(1-y^2) = 2y$ with $y = \tanh t \in (-1,1)$. Solving $cy^2 + 2y - c = 0$ yields the feasible root $y^\star = \frac{\sqrt{1+c^2}-1}{c} = \frac{c}{1+\sqrt{1+c^2}}$. At $y^\star$, $\operatorname{sech}^2 t = 1-(y^\star)^2 = \frac{2}{1+\sqrt{1+c^2}}$ and $c\tanh t = \frac{c^2}{1+\sqrt{1+c^2}}$. Substituting into $f(t)$ gives Eq. (27). □

B.5 Certified Per-Token Gradient Control and Smoothness Guarantees

This subsection provides the full statements and proof details for the certified moment-control chain of Theorem 2. We keep the assumptions explicit and local to one policy update, which is the regime in which PPO/GRPO-style policy-gradient estimators are applied.

Assumptions. Fix the current parameter value and behavior law μ. The main theorem assumes Eq. (18), retaining any dependence between advantages, score norms, and response lengths. For the auxiliary per-token bounds below, stronger sufficient conditions are

$$\mathbb{E}_\mu[A_t^2] \le \sigma_A^2, \tag{28}$$

$$\|\nabla_\theta \log \pi_\theta(y_t \mid x, y_{<t})\| \le G \qquad \mu\text{-a.s.} \tag{29}$$

Eq. (29) is an almost-sure per-token score assumption. For a fixed token, these stronger assumptions imply $\mathbb{E}[A_t^2\|s_t\|^2] \le \sigma_A^2 G^2$ by pointwise domination. For the random-length mini-batch, Eq. (18) is stated directly on the normalized sum at the current parameter value.

For curvature discussions one may additionally assume, along the line segment of the realized update, that

$$\left\|\nabla_{\tilde\theta}^2 \log \pi_{\tilde\theta}(y_t \mid x, y_{<t})\right\| \le H, \tag{30}$$

where $\|\cdot\|$ denotes the Euclidean norm for vectors and the induced operator norm for matrices. This curvature condition is not needed for Theorem 2; it only supports smoothness interpretations of the LFM path.

For the scalar LFM log-operator $u(z) = c\tanh(z/c)$, Lemma 7 gives $0 < u'(z) \le 1$ and $|u''(z)| \le 2/c$. For MHPO under the semi-gradient implementation, the effective derivative is $u'(z)$ because Ω is stop-gradient, while the operator value is $\exp(u(z) - \Omega(z)) \le \exp(u(z)) \le e^c$ by construction. This operator-value interval is distinct from the sharper multiplier certificate $M_\psi(c)$ used for second-moment and tail control.

Lemma 9 (Certified operator-value interval)**.** *For $u(z) = c\tanh(z/c)$, we have $|u(z)| \le c$ and hence $\exp(u(z)) \in [e^{-c}, e^c]$ for all $z \in \mathbb{R}$. In* MHPO*, since the hazards are nonnegative, $\Omega(z) \ge 0$ and thus $\omega(e^z) = \exp(u(z) - \Omega(z)) \le \exp(u(z)) \le e^c$.*

Proof. Since $|\tanh(t)| \le 1$ for all $t \in \mathbb{R}$, we have

$$|u(z)| = c\left|\tanh(z/c)\right| \le c.$$

Thus $-c \le u(z) \le c$, and exponentiating yields $e^{-c} \le \exp(u(z)) \le e^c$ for all z. For MHPO, since $\Omega(z) \ge 0$, we have $\omega(e^z) = \exp(u(z) - \Omega(z)) \le \exp(u(z)) \le e^c$. This bounds the forward ratio operator value; the gradient multiplier is additionally scaled by $u'(z)$ and is therefore controlled by the exact supremum $M_\psi(c)$ in Theorem 8. □

Proposition 10 (Bounded gradient multiplier implies per-token gradient control)**.** *Let $\ell_t(\theta) = -A_t \exp(\phi(z_t(\theta)))$ and $\mathcal{M}_\phi(z) = \exp(\phi(z))\phi'(z)$. Under Assumption (29),*

$$\|\nabla_\theta \ell_t(\theta)\| \le |A_t| \sup_z |\mathcal{M}_\phi(z)|\, G.$$

Moreover, if $\mathbb{E}[A_t^2] \le \sigma_A^2$, then

$$\mathbb{E}\Big[\|\nabla_\theta \ell_t(\theta)\|^2\Big] \le G^2\, \mathbb{E}\big[A_t^2\, \mathcal{M}_\phi(z_t(\theta))^2\big] \le \sigma_A^2\, G^2 \left(\sup_z |\mathcal{M}_\phi(z)|\right)^2.$$

Proof. Lemma 4 gives the chain-rule form

$$\nabla_\theta \ell_t(\theta) = -A_t \, \exp(\phi(z_t(\theta))) \, \phi'(z_t(\theta)) \, \nabla_\theta z_t(\theta) = -A_t \, \mathcal{M}_\phi(z_t(\theta)) \, \nabla_\theta z_t(\theta),$$

where $\mathcal{M}_\phi(z) \triangleq \exp(\phi(z))\phi'(z)$. By Lemma 5, $\nabla_\theta z_t(\theta) = \nabla_\theta \log \pi_\theta(y_t \mid x, y_{<t})$. Substituting yields

$$\nabla_\theta \ell_t(\theta) = -A_t \, \mathcal{M}_\phi(z_t(\theta)) \, \nabla_\theta \log \pi_\theta(y_t \mid x, y_{<t}).$$

Taking norms and using Assumption (29) gives

$$\begin{aligned} \|\nabla_\theta \ell_t(\theta)\| &\le |A_t| \, |\mathcal{M}_\phi(z_t(\theta))| \, \big\|\nabla_\theta \log \pi_\theta(y_t \mid x, y_{<t})\big\| \\ &\le |A_t| \, |\mathcal{M}_\phi(z_t(\theta))| \, G \\ &\le |A_t| \, \Big(\sup_{z\in\mathbb{R}} |\mathcal{M}_\phi(z)|\Big) \, G, \end{aligned}$$

which proves the first inequality.

Squaring the exact expression and applying $\|\nabla_\theta \log \pi_\theta(\cdot)\| \le G$ yields

$$\|\nabla_\theta \ell_t(\theta)\|^2 = A_t^2 \, \mathcal{M}_\phi(z_t(\theta))^2 \, \big\|\nabla_\theta \log \pi_\theta(y_t \mid x, y_{<t})\big\|^2 \le G^2 \, A_t^2 \, \mathcal{M}_\phi(z_t(\theta))^2.$$

Taking expectations gives

$$\begin{aligned} \mathbb{E}\big[\|\nabla_\theta \ell_t(\theta)\|^2\big] &\le G^2 \, \mathbb{E}\big[A_t^2 \, \mathcal{M}_\phi(z_t(\theta))^2\big] \\ &\le G^2 \, \Big(\sup_{z\in\mathbb{R}} |\mathcal{M}_\phi(z)|\Big)^2 \, \mathbb{E}[A_t^2] \\ &\le \sigma_A^2 \, G^2 \, \Big(\sup_{z\in\mathbb{R}} |\mathcal{M}_\phi(z)|\Big)^2, \end{aligned}$$

which proves the second-moment bound. □

Lemma 11 (Markov's inequality)**.** *Let X be a nonnegative random variable with $\mathbb{E}[X] < \infty$. Then for any $\tau > 0$,*

$$\Pr(X \ge \tau) \le \frac{\mathbb{E}[X]}{\tau}.$$

Proof. Since $X \ge 0$, we have $X\mathbf{1}\{X \ge \tau\} \ge \tau \, \mathbf{1}\{X \ge \tau\}$. Taking expectations yields

$$\mathbb{E}[X] \ge \mathbb{E}\big[X\mathbf{1}\{X \ge \tau\}\big] \ge \tau \, \Pr(X \ge \tau),$$

which rearranges to the claimed bound. □

Corollary 12 (Tail bound from a second moment)**.** *Let Z be a (possibly vector-valued) random variable with $\mathbb{E}[\|Z\|^2] < \infty$. Then for any $\tau > 0$,*

$$\Pr(\|Z\| \ge \tau) = \Pr(\|Z\|^2 \ge \tau^2) \le \frac{\mathbb{E}[\|Z\|^2]}{\tau^2}.$$

Proof. Apply Lemma 11 to the nonnegative random variable $X = \|Z\|^2$ with threshold τ^2. □

Proof of Theorem 2. Because $\Omega(z) \ge 0$ and $u_c'(z) = \operatorname{sech}^2(z/c) > 0$, the semi-gradient multiplier satisfies

$$0 \le \mathcal{M}(z) = \exp(u_c(z) - \Omega(z))u_c'(z) \le \exp(u_c(z))u_c'(z).$$

Taking the supremum over z and applying Theorem 8 gives the multiplier bound in Eq. (19).

For a token in response i, define

$$v_t^i = \mathcal{M}(z_t^i(\theta)) \, \nabla_\theta \log \pi_\theta(q_t^i \mid p, q_{<t}^i) \, \hat{A}_t^i.$$

By the pointwise multiplier bound,

$$\|v_t^i\|^2 \le M_\psi(c)^2 (\hat{A}_t^i)^2 \|s_t^i\|^2.$$

For each realized mini-batch, let $h_i = (1/T_i)\sum_{t=1}^{T_i} v_t^i$. Convexity of $\|\cdot\|^2$ gives

$$\|h_i\|^2 = \left\| \frac{1}{T_i} \sum_{t=1}^{T_i} v_t^i \right\|^2 \le \frac{1}{T_i} \sum_{t=1}^{T_i} \|v_t^i\|^2.$$

Applying the same convexity argument to $g(\theta) = (1/K)\sum_i h_i$ gives

$$\|g(\theta)\|^2 \le \frac{1}{K} \sum_{i=1}^{K} \|h_i\|^2 \le \frac{1}{K} \sum_{i=1}^{K} \frac{1}{T_i} \sum_{t=1}^{T_i} \|v_t^i\|^2.$$

Substituting the pointwise bound and taking expectations of the entire normalized sum yields $\mathbb{E}_\mu\|g\|^2 \le M_\psi(c)^2 C$ by Eq. (18). This handles random lengths without assuming that unconditional token moments also bound length-conditional moments. Corollary 12 gives the tail bound. No token independence is used. The pointwise multiplier factor remains valid under any behavior law, but its associated product-moment constant must be established for that law. □

B.6 Mismatch Robustness and the Rationale for Asymmetric Accelerating Hazards

We next state three results connecting the operator to changed behavior laws, autoregressive branch contraction, and accelerating tail regularization.

Proposition 13 (Multiplier control under a changed behavior law)**.** *Let μ be any behavior distribution for which the per-token semi-gradient*

$$g_t(\theta) \triangleq -A_t\, \mathcal{M}(z_t(\theta))\, \nabla_\theta \log \pi_\theta(a_t \mid s_t)$$

is well-defined, where $z_t(\theta) = \log[\pi_\theta(a_t \mid s_t)/\pi_{\theta_{\text{old}}}(a_t \mid s_t)]$ and $\mathcal{M}(z) = e^{u(z)-\Omega(z)}u'(z)$ is the implemented semi-gradient multiplier. Assume $\mathbb{E}_\mu[A_t^2] \le \sigma_A^2$ and $\|\nabla_\theta \log \pi_\theta(a_t \mid s_t)\| \le G$ almost surely. Then

$$\mathbb{E}_\mu\big[\|g_t(\theta)\|^2\big] \le \sigma_A^2 G^2\, M_\psi(c)^2.$$

The same bound holds for deterministic convex averages of these gradients. For random response lengths and weights, the normalized product-moment condition in Eq. (18) gives the main theorem's bound instead.

Proof. By Proposition 10 and Theorem 8,

$$\|g_t(\theta)\| \le |A_t| \sup_z |\mathcal{M}(z)|\, G \le |A_t|\, M_\psi(c)\, G.$$

Squaring and taking expectation under μ gives

$$\mathbb{E}_\mu\big[\|g_t(\theta)\|^2\big] \le M_\psi(c)^2 G^2\, \mathbb{E}_\mu[A_t^2] \le \sigma_A^2 G^2\, M_\psi(c)^2.$$

If $\bar g = \sum_{j=1}^N \alpha_j g_j$ with deterministic $\alpha_j \ge 0$ and $\sum_j \alpha_j = 1$, then convexity of $\|\cdot\|^2$ yields

$$\|\bar g\|^2 \le \sum_{j=1}^{N} \alpha_j \|g_j\|^2.$$

Taking expectation and applying the per-token bound to each g_j proves the mini-batch claim. □

Proposition 13 separates the pointwise ratio-multiplier bound from the advantage–score moments. The multiplier bound holds under each behavior law, while the corresponding gradient moment bound uses the moments under that law.

Proposition 14 (Ratio contractions contribute multiplicatively to prefix occupancy)**.** *Let π and μ be two autoregressive policies on the same support. For any realized prompt–trajectory pair (x, y) and any prefix length $t \ge 1$, define the prefix occupancies*

$$d_\pi(x, y_{<t}) = P(x) \prod_{j=0}^{t-1} \pi(y_j \mid x, y_{<j}), \qquad d_\mu(x, y_{<t}) = P(x) \prod_{j=0}^{t-1} \mu(y_j \mid x, y_{<j}),$$

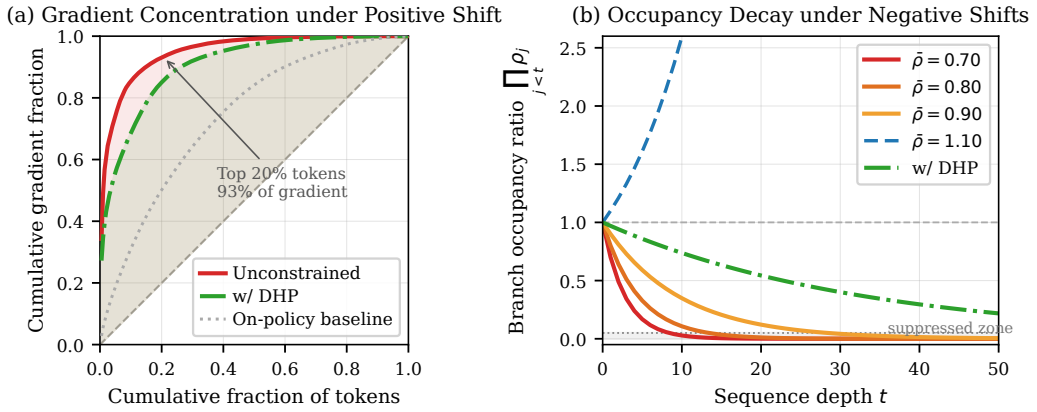


Figure 4: **Gradient concentration and prefix contraction. (a)** Cumulative gradient contribution sorted in descending order measures the concentration of the learning signal across tokens. **(b)** Prefix likelihood ratios $R_t = \bar{\rho}^t$ for constant token ratios $\bar{\rho} = 0.90$ and 0.97. At 50 tokens, the ratios are 0.0052 and 0.218, respectively. Both contract relative to the behavior policy, with slower contraction for $\bar{\rho} = 0.97$. These curves evaluate the fixed-ratio product identity.

and the token ratios $\rho_j = \frac{\pi(y_j|x,y_{<j})}{\mu(y_j|x,y_{<j})}$ *with* $z_j = \log \rho_j$. *Then*

$$\frac{d_\pi(x, y_{<t})}{d_\mu(x, y_{<t})} = \prod_{j=0}^{t-1} \rho_j = \exp\left(\sum_{j=0}^{t-1} z_j\right).$$

Consequently, if a subset $J \subseteq \{0, \ldots, t-1\}$ *satisfies* $z_j \leq -\delta_j < 0$ *for all* $j \in J$, *then every descendant prefix containing those contractions obeys*

$$\frac{d_\pi(x, y_{<t})}{d_\mu(x, y_{<t})} \leq \exp\left(-\sum_{j \in J} \delta_j\right) \prod_{k \notin J,\, k<t} \rho_k.$$

In particular, n *contractions of magnitude at least* δ *contribute a factor at most* $e^{-n\delta}$ *relative to the product of the remaining ratios. The full occupancy ratio combines this contraction factor with the remaining factors, including expansions.*

Proof. Substituting the definitions of d_π and d_μ gives

$$\frac{d_\pi(x, y_{<t})}{d_\mu(x, y_{<t})} = \frac{P(x)\prod_{j=0}^{t-1} \pi(y_j \mid x, y_{<j})}{P(x)\prod_{j=0}^{t-1} \mu(y_j \mid x, y_{<j})} = \prod_{j=0}^{t-1} \frac{\pi(y_j \mid x, y_{<j})}{\mu(y_j \mid x, y_{<j})} = \prod_{j=0}^{t-1} \rho_j.$$

Writing $\rho_j = e^{z_j}$ yields the exponential form. If $z_j \leq -\delta_j$ for $j \in J$, then

$$\prod_{j=0}^{t-1} \rho_j = \Big(\prod_{j \in J} e^{z_j}\Big)\Big(\prod_{k \notin J,\, k<t} \rho_k\Big) \leq \exp\left(-\sum_{j \in J} \delta_j\right) \prod_{k \notin J,\, k<t} \rho_k.$$

Setting $\delta_j \geq \delta$ for $n = |J|$ gives the final bound. □

Proposition 14 identifies how repeated contractions reduce the likelihood of a fixed branch relative to its other ratio factors. This motivates protecting useful prefixes against cumulative contraction. Section C connects the identity to local updates and empirical prefix diagnostics.

Proposition 15 ($k > 1$ is the accelerating-hazard regime)**.** *Let* $H(x) = (x/\lambda)^k$ *be the Weibull cumulative hazard, where* $\lambda > 0$ *and* $x > 0$. *Then*

$$\frac{H(x)}{x} = \frac{x^{k-1}}{\lambda^k}.$$

If $k > 1$, this ratio is strictly increasing, with

$$\lim_{x \downarrow 0} \frac{H(x)}{x} = 0, \qquad \lim_{x \to \infty} \frac{H(x)}{x} = \infty.$$

Hence for any $0 < \eta_{\text{in}} < \eta_{\text{out}}$, there exist $0 < x_{\text{in}} < x_{\text{out}} < \infty$ such that

$$H(x) \le \eta_{\text{in}} x \quad \textit{for all } x \in [0, x_{\text{in}}], \qquad H(x) \ge \eta_{\text{out}} x \quad \textit{for all } x \ge x_{\text{out}}.$$

If $k = 1$, then $H(x)/x \equiv 1/\lambda$ is constant, so no such scale separation is possible.

Proof. The identity $H(x)/x = x^{k-1}/\lambda^k$ is immediate. When $k > 1$, its derivative

$$\frac{d}{dx}\Big(\frac{H(x)}{x}\Big) = \frac{k-1}{\lambda^k} x^{k-2}$$

is strictly positive for $x > 0$, so the ratio is strictly increasing. The two limits follow directly. Therefore, given $0 < \eta_{\text{in}} < \eta_{\text{out}}$, we may choose x_{in} small enough that $H(x)/x \le \eta_{\text{in}}$ on $(0, x_{\text{in}}]$ and x_{out} large enough that $H(x)/x \ge \eta_{\text{out}}$ on $[x_{\text{out}}, \infty)$. When $k = 1$, the ratio is the constant $1/\lambda$, so it cannot be simultaneously arbitrarily small near the origin and arbitrarily large in the tail. □

Proposition 15 describes the Weibull family on the domain $x > 0$. In MHPO, $x_{\pm}(u(z))$ remain in a compact interval because LFM bounds u. The operational property is therefore the increase of $H(x)/x$ over the attained magnitudes. Also, $\partial H/\partial k = H \log(x/\lambda)$, so increasing k strengthens attenuation for $x > \lambda$ and weakens it for $x < \lambda$. This gives separate control over attenuation scale and central-to-tail contrast.

Connection to the design. The three propositions explain complementary design choices. Proposition 13 bounds ratio-induced amplification under the stated moment conditions. Proposition 14 identifies the cumulative effect of contraction on branch likelihood. Proposition 15 explains how Weibull shaping controls the increase in attenuation across directional magnitudes.

B.7 Full-Gradient Analysis of Hazard Shaping and Failure Modes

This final subsection provides the formal justification for the semi-gradient (stop-gradient) implementation of hazard shaping adopted in the main paper. After establishing the bounded semi-gradient analysis above, we contrast it with the fully differentiable (no stop-gradient) variant, which can reverse the gradient sign and does not inherit the same c-only attenuation certificate.

B.7.1 Full-Gradient Derivation of $\Omega'(z)$

Let $u(z) = c \cdot \tanh(z/c)$ (equivalently, $u(z) = \psi(e^z)$), so that $u'(z) = \operatorname{sech}^2(z/c)$. The one-sided magnitudes are $x_+(u) = \text{softplus}(\beta u)/\beta$ and $x_-(u) = \text{softplus}(-\beta u)/\beta$, with derivatives

$$\frac{dx_+}{du} = \text{sigmoid}(\beta u), \qquad \frac{dx_-}{du} = -\,\text{sigmoid}(-\beta u) = -(1 - \text{sigmoid}(\beta u)). \tag{31}$$

The cumulative Weibull hazard $H_+(x_+) = (x_+/\lambda_+)^{k_+}$ has derivative

$$\frac{dH_+}{dx_+} = \frac{k_+}{\lambda_+}\left(\frac{x_+}{\lambda_+}\right)^{k_+ - 1}. \tag{32}$$

Applying the chain rule through $H_+ \circ x_+ \circ u$ and $H_- \circ x_- \circ u$, the full derivative of $\Omega(z) = H_+(x_+(u(z))) + H_-(x_-(u(z)))$ is

$$\Omega'(z) = \operatorname{sech}^2\Big(\frac{z}{c}\Big)\left[\frac{k_+}{\lambda_+}\left(\frac{x_+}{\lambda_+}\right)^{k_+ - 1}\text{sigmoid}(\beta u) - \frac{k_-}{\lambda_-}\left(\frac{x_-}{\lambda_-}\right)^{k_- - 1}\big(1 - \text{sigmoid}(\beta u)\big)\right]. \tag{33}$$

B.7.2 Full-Gradient Multiplier Analysis and Sign-Reversal Phenomenon

Without stop-gradient, the log-form $\phi(z) = u(z) - \Omega(z)$ is fully differentiable, and the gradient multiplier becomes

$$\mathcal{M}_{\text{full}}(z) = \exp\big(\phi(z)\big) \cdot \phi'(z) = \exp\big(u(z) - \Omega(z)\big) \cdot \big[u'(z) - \Omega'(z)\big]. \tag{34}$$

The critical term is $u'(z) - \Omega'(z)$, which determines the sign of the gradient multiplier. Substituting and factoring out $\operatorname{sech}^2(z/c)$, we obtain

$$\phi'(z) = \operatorname{sech}^2\Big(\frac{z}{c}\Big) \underbrace{\left[1 - \frac{k_+}{\lambda_+}\left(\frac{x_+}{\lambda_+}\right)^{k_+ - 1} \operatorname{sigmoid}(\beta u) + \frac{k_-}{\lambda_-}\left(\frac{x_-}{\lambda_-}\right)^{k_- - 1} \big(1 - \operatorname{sigmoid}(\beta u)\big)\right]}_{\triangleq\ D(z)}. \tag{35}$$

Since $\operatorname{sech}^2(z/c) > 0$ everywhere, the sign of $\mathcal{M}_{\text{full}}(z)$ is entirely determined by $D(z)$.

Problem 1: Parameter-dependent gradient sign reversal. LFM bounds $u(z)$ inside $[-c, c]$, keeping the composed hazard derivative bounded. The full-gradient multiplier changes sign wherever $\Omega'(z) > u'(z)$, which is determined by the directional parameters after removing the common factor $\operatorname{sech}^2(z/c)$. Define, for $a \in [-c, c]$,

$$D_c(a) = 1 - \frac{k_+}{\lambda_+}\left(\frac{x_+(a)}{\lambda_+}\right)^{k_+ - 1} \operatorname{sigmoid}(\beta a) + \frac{k_-}{\lambda_-}\left(\frac{x_-(a)}{\lambda_-}\right)^{k_- - 1} \big(1 - \operatorname{sigmoid}(\beta a)\big).$$

Since $u(z) \to c$ as $z \to \infty$ and $u(z) \to -c$ as $z \to -\infty$, we have $\lim_{z\to\infty} D(z) = D_c(c)$ and $\lim_{z\to-\infty} D(z) = D_c(-c)$. Therefore, if $D_c(c) < 0$, there exists a positive-tail region in which $\mathcal{M}_{\text{full}}(z) < 0$; if $D_c(-c) < 0$, the analogous negative-tail sign flip occurs. For such configurations, a token with positive advantage can receive a reversed gradient contribution, because

$$\nabla_\theta \ell_t = -\hat{A}_t\, \mathcal{M}_{\text{full}}(z_t)\, \nabla_\theta \log \pi_\theta(\cdot)$$

changes direction when $\mathcal{M}_{\text{full}}(z_t)$ becomes negative. The semi-gradient implementation removes this failure mode by replacing $\phi'(z) = u'(z) - \Omega'(z)$ with the nonnegative effective derivative $u'(z)$.

Problem 2: Loss of the c-only attenuation certificate. Under the semi-gradient implementation, the gradient multiplier becomes $\mathcal{M}_{\text{sg}}(z) = \exp(u(z) - \Omega(z))\, u'(z) = \exp(u(z))\, w(z)\, u'(z)$ with $w(z) = \exp(-\Omega(z)) \in (0, 1]$. Therefore, its peak is bounded by the exact LFM base certificate $M \le M_\psi(c)$ (Theorem 8). In the full-gradient case, bounding $|\mathcal{M}_{\text{full}}(z)|$ requires analyzing $|D(z)|$, which simultaneously depends on the Weibull parameters $(k_\pm, \lambda_\pm)$, the softplus smoothing β, and the bound c. The branch magnitudes are bounded away from zero and infinity, so $D_c(a)$ is continuous on $[-c, c]$ and yields the finite bound $|\mathcal{M}_{\text{full}}(z)| \le M_\psi(c) \max_{a\in[-c,c]} |D_c(a)|$. Nonnegativity additionally requires $D_c(a) \ge 0$ throughout the attained interval. The implemented semi-gradient removes this parameter-dependent factor, preserving the c-only bound and the direction of each direct contribution.

C Mechanism: Retained Signals, Directional Tails, and Prefix Dynamics

MHPO retains useful policy-gradient signals while limiting the influence of extreme ratio deviations. The analysis connects three effects: LFM restores useful clipped-tail contributions, DHP selectively reduces contraction tails, and these directional changes accompany improved retention of correct-response prefixes. Token interventions and directional diagnostics connect the operators to their observed effects.

C.1 From GRPO Clipping to Retained Token Contributions

Fix a sampled token, its advantage $\hat{A}_t$, and the behavior policy. Let $z_t = \log r_t$ and $s_t = \nabla_\theta \log \pi_\theta(a_t \mid s_t^{\text{state}})$. The symbol s_t^{state} denotes the conditioning prefix, while the vector s_t denotes the score. For the maximization surrogate,

$$J_{\text{GRPO},t} = \min\{r_t \hat{A}_t, \operatorname{clip}(r_t, 1-\epsilon, 1+\epsilon)\hat{A}_t\}, \qquad \nabla_\theta r_t = r_t s_t. \tag{36}$$

When $\hat{A}_t > 0$, the upper branch becomes constant for $r_t > 1 + \epsilon$. When $\hat{A}_t < 0$, the lower branch becomes constant for $r_t < 1 - \epsilon$. Away from the boundaries,

$$D_t = \mathbf{1}[(\hat{A}_t > 0 \wedge r_t > 1 + \epsilon) \vee (\hat{A}_t < 0 \wedge r_t < 1 - \epsilon)], \tag{37}$$

$$\nabla_\theta J_{\mathrm{GRPO},t} = \hat{A}_t r_t (1 - D_t) s_t. \tag{38}$$

Zero-advantage tokens contribute nothing regardless of clipping and are excluded from the dead-token diagnostics. A nonzero-advantage token outside the clip interval is not necessarily dead. For instance, $\hat{A}_t < 0$ and $r_t > 1$ gives an active corrective contribution that decreases the token probability locally.

LFM instead uses $u(z) = c \tanh(z/c)$ and

$$\mathcal{M}_{\mathrm{LFM}}(z) = e^{u(z)} u'(z) = e^{c \tanh(z/c)} \mathrm{sech}^2(z/c), \qquad \nabla_\theta J_{\mathrm{LFM},t} = \hat{A}_t \mathcal{M}_{\mathrm{LFM}}(z_t) s_t. \tag{39}$$

Its multiplier is positive for finite z, locally agrees with $r = e^z$, and decays in both tails. Hard clipping switches some contributions off, whereas LFM attenuates them continuously. The following diagnostics quantify their retained mass, update alignment, and contribution to performance.

For a diagnostic token set $\mathcal{I}$ with $\hat{A}_t \neq 0$, define

$$f_D = \frac{\sum_{t \in \mathcal{I}} D_t}{|\mathcal{I}|}, \qquad m_D = \frac{\sum_{t \in \mathcal{I}} D_t |\hat{A}_t|}{\sum_{t \in \mathcal{I}} |\hat{A}_t|}, \tag{40}$$

$$m_D^{\mathrm{LFM}} = \frac{\sum_{t \in \mathcal{I}} D_t |\hat{A}_t| \mathcal{M}_{\mathrm{LFM}}(z_t)}{\sum_{t \in \mathcal{I}} |\hat{A}_t| \mathcal{M}_{\mathrm{LFM}}(z_t)}, \tag{41}$$

$$\Delta_{\mathrm{align},t} = \mathrm{sign}(\hat{A}_t) \left[\log \pi_{\mathrm{after}}(a_t \mid s_t^{\mathrm{state}}) - \log \pi_{\mathrm{before}}(a_t \mid s_t^{\mathrm{state}}) \right]. \tag{42}$$

Here f_D counts tokens, m_D weights them by advantage magnitude, and m_D^{LFM} weights them by LFM's effective scalar coefficient. Coefficient mass is not parameter-gradient norm mass, because score-vector norms and directions also matter. Positive Δ_{align} means the realized log-probability change agrees with the advantage direction. GRPO can have nonzero realized changes on its dead set through other tokens' gradients, parameter sharing, and optimizer state.

Signal-retention evidence. At late checkpoints, $f_D = 18.6\%$, $m_D = 27.4\%$, and $m_D^{\mathrm{LFM}} = 34.7\%$. On this set, the average advantage-aligned log-probability change is 0.002 with GRPO and 0.019 with LFM. LFM's best/latest Avg@32 is $52.0/48.5$. Applying the exact GRPO-tail mask to its contributions lowers these scores to $49.2/42.6$. A shuffled mask matched in count, position, advantage sign, $|A|$ bin, and removed coefficient mass gives $51.1/46.8$. The exact mask is more damaging than removing comparable mass elsewhere, showing that the recovered clipped-tail signals contribute to task performance and checkpoint retention.

C.2 What DHP Does to a Local Update

Under the implemented stop-gradient,

$$\mathcal{M}_{\mathrm{MHPO}}(z) = w(z) \mathcal{M}_{\mathrm{LFM}}(z), \qquad w(z) = e^{-\Omega(z)} \in (0, 1]. \tag{43}$$

This is an exact identity for a direct token contribution. To connect it to a probability change, consider one isolated-token SGD ascent step at the same initial θ, with fixed $\hat{A}_t$, s_t, and learning rate η:

$$\delta\theta_{\mathrm{LFM}} = \eta \hat{A}_t \mathcal{M}_{\mathrm{LFM}}(z_t) s_t, \tag{44}$$

$$\delta\theta_{\mathrm{MHPO}} = \eta \hat{A}_t \mathcal{M}_{\mathrm{MHPO}}(z_t) s_t = w(z_t) \delta\theta_{\mathrm{LFM}}. \tag{45}$$

$\delta\theta_{\mathrm{LFM}}$ denotes the parameter displacement caused by this token under LFM alone, evaluated at the common initial θ.

Let $\Delta z_X = z_t(\theta + \delta\theta_X) - z_t(\theta)$ for $X \in \{\mathrm{LFM}, \mathrm{MHPO}\}$. If the score Hessian has norm at most H along the two step segments, Taylor's theorem gives

$$\Delta z_{\mathrm{LFM}} = s_t^\top \delta\theta_{\mathrm{LFM}} + R_L, \quad |R_L| \leq \tfrac{H}{2} \|\delta\theta_{\mathrm{LFM}}\|^2, \tag{46}$$

$$\Delta z_{\mathrm{MHPO}} = w(z_t) s_t^\top \delta\theta_{\mathrm{LFM}} + R_M, \quad |R_M| \leq \tfrac{H}{2} w(z_t)^2 \|\delta\theta_{\mathrm{LFM}}\|^2. \tag{47}$$

Subtracting $w(z_t)\Delta z_{\mathrm{LFM}}$ yields

$$|\Delta z_{\mathrm{MHPO}} - w(z_t)\Delta z_{\mathrm{LFM}}| \le \frac{H}{2}\big(w(z_t)^2 + w(z_t)\big)\|\delta\theta_{\mathrm{LFM}}\|^2 = O(\eta^2). \tag{48}$$

The parameter-step relation is exact in this isolated-token setup, whereas the log-probability relation is first-order. Since $s_t^\top \delta\theta_{\mathrm{LFM}} = \eta\hat{A}_t\mathcal{M}_{\mathrm{LFM}}\|s_t\|^2$, a negative advantage causes a direct decrease. If $z_t < 0$ already, this reinforces an existing contraction, and DHP reduces its first-order magnitude. If $\hat{A}_t > 0$ with $z_t < 0$, the update is corrective, and DHP also attenuates that correction. Thus ratio direction controls attenuation, while advantage sign determines the direct update direction.

Shared-parameter updates. For a mini-batch with normalized weights α_t, the first-order SGD expansion underlying Eq. (12) gives the evaluated token j

$$\Delta z_j = \eta \sum_t \alpha_t \hat{A}_t \mathcal{M}_t \, s_j^\top s_t + O(\eta^2). \tag{49}$$

Cross-token inner products can have either sign, and DHP applies a token-specific w_t to each term. Consequently, mini-batch updates combine differently weighted score interactions. Optimizer state and repeated actor steps further shape the realized probability changes. The directional-tail and prefix diagnostics below measure these effects during training.

C.3 Directional-Tail Severity and CVaR

Negative-tail 1%-CVaR (conditional value at risk) is the mean contraction severity in the worst 1% of tokens. Define $L = (-z)_+ = \max(0, -\log r)$ and tail fraction $\alpha = 0.01$. A definition valid even when the distribution has atoms is

$$\mathrm{CVaR}_\alpha^{\mathrm{upper}}(L) = \min_{v\in\mathbb{R}} \left\{ v + \frac{1}{\alpha}\mathbb{E}[(L-v)_+] \right\}. \tag{50}$$

For a continuous distribution, this equals $\mathbb{E}[L \mid L \ge Q_{1-\alpha}(L)]$. For an empirical sample, replace the expectation with its sample mean. This averages the top α fraction, with fractional boundary weight if necessary. VaR identifies where the worst tail starts. CVaR says how severe that tail is on average. Because larger L means a smaller token ratio, a lower value indicates less severe contraction. CVaR 1.18 corresponds to a geometric mean ratio $e^{-1.18} \approx 0.307$ within that negative tail, not to an arithmetic mean ratio or a prefix probability.

Tail frequency and severity measure different properties. With GRPO's $\epsilon = 0.2$, define

$$f_- = \Pr(z < \log(1-\epsilon)), \qquad f_+ = \Pr(z > \log(1+\epsilon)). \tag{51}$$

Unlike the dead set, these quantities describe ratio direction without conditioning on advantage sign. The negative-tail frequency and CVaR are 8.9% and 1.18 with LFM, versus 3.9% and 0.63 with MHPO. A per-batch scalar attenuation control matched to MHPO's update norm gives 8.0% and 1.09. MHPO therefore reduces contraction frequency and severity beyond the uniform attenuation control. Compared with symmetric DHP, MHPO retains more expansion-side variation, with $f_+ = 7.1\%$ versus 5.8% and a highest-1% mean z of 0.91 versus 0.77. Together, these comparisons show selective regulation of the two directional tails.

C.4 From Token Contraction to Prefix Likelihood

For a fixed response with a common prompt and behavior-policy reference,

$$C_{i,t} = \sum_{j=1}^{t} \log r_{i,j} = \log \frac{\pi_\theta(q^i_{\le t} \mid p)}{\pi_{\theta_{\mathrm{old}}}(q^i_{\le t} \mid p)}, \qquad R_{i,t} = e^{C_{i,t}}. \tag{52}$$

Each token ratio is a multiplicative factor in $R_{i,t}$. For example, ten factors of 0.9 contribute $0.9^{10} \approx 0.349$ relative to the remaining factors. The full prefix ratio combines these contractions with the expansions along the same response, motivating a trajectory-level measure of their cumulative effect.

For threshold $a = 0.1$, define

$$F_a(i) = \mathbf{1}\left[\min_{1\le t\le T_i} C_{i,t} < \log a\right], \qquad \tau_a(i) = \inf\{t : C_{i,t} < \log a\}. \tag{53}$$

The collapse rate is the fraction of responses with $F_a(i) = 1$. It means that at some depth the prefix is assigned less than one tenth of its behavior-policy likelihood, not that the final answer is incorrect. For noncrossing responses, $\tau_a = \infty$. The reported first-crossing median is conditional on $F_a = 1$, excluding noncrossers rather than assigning them the maximum response length. Because the crossing subset can differ between methods, the conditional median should be interpreted alongside the collapse rate, not as an unconditional delay estimate.

On correct responses, the collapse rate is 13.2% for LFM, 11.5% for the norm-matched scalar control, and 4.9% for MHPO. Among crossing responses, the median first-crossing depth changes from 846 tokens with LFM to $1{,}487$ with MHPO. Correct responses directly expose the loss of useful trajectory probability. Their lower collapse rate connects the observed directional-tail reduction to improved prefix retention.

Connection back to the motivation. The evidence follows the design's two roles. LFM retains nonnegligible, advantage-aligned coefficients in regions that GRPO clips away, and masking those specific contributions harms performance. DHP changes contraction-tail frequency and severity beyond the tested scalar control, accompanied by fewer collapsed correct-response prefixes. The component results (50.8 for DHP, 52.0 for LFM, and 52.9 for the combination) and checkpoint retention in Table 6 are consistent with this division of labor. Together, the results support useful signal retention and selective contraction control as complementary mechanisms underlying MHPO's performance.

D Additional Results Across Backbone Settings

This appendix provides additional main-result evaluations on instruction-tuned backbones, which complement the foundation-model results reported in the main paper on Qwen3-4B-Base and Qwen3-30B-A3B-Base. Table 5 reports Avg@32 on three instruction-tuned backbones: a general-purpose model (Qwen2.5-7B-Instruct), a domain-specialized model (Qwen2.5-Math-7B-Instruct), and a vision-language model (Qwen2.5-VL-7B-Instruct).

Domain-specialized and general-purpose models. On the general-purpose Qwen2.5-7B-Instruct backbone (Table 5A), MHPO reaches a macro-average of 43.7%, outperforming the strongest baseline DAPO by 3.5 points. On the domain-specialized Qwen2.5-Math-7B-Instruct (Table 5B), MHPO reaches 47.8%, exceeding DAPO by 4.2 points. The gains are largest on the most challenging benchmarks. On HMMT25, MHPO reaches 17.9% against DAPO's 11.5%, a gap of $+6.4$ points. On AIME25, MHPO reaches 21.6% against 15.0%, a gap of $+6.6$ points. These results indicate that the proposed framework delivers consistent improvements even when the backbone already carries strong domain priors.

Text-only and vision-language models. On Qwen2.5-VL-7B-Instruct (Table 5C), MHPO reaches 53.0%, 5.5 points above the initial backbone and 0.7 points above DAPO. The MathVision gains are 5.3 and 1.4 points, respectively. These results extend the observed improvements from text-only to vision-language reasoning.

The complementary distilled-model comparison appears in Section 4.4 and Table 4. Across these Qwen-derived backbones, MHPO uses the same operator interface of token log-probabilities and advantages.

E Checkpoint Robustness Analysis

This appendix provides the full per-benchmark comparison between the best (peak) and latest (final) checkpoints on Qwen3-4B-Base. Table 6 expands the summary statement given in the training-stability analysis of the main paper.

DAPO's macro-average drops from 43.6% to 31.4%, while the other baselines lose 6.9–8.2 points. MHPO loses 0.8 points, from 52.9% to 52.1%. This improved checkpoint retention is consistent with the token and prefix diagnostics in Section C: LFM retains useful learning signals, DHP reduces directional contraction, and their combination retains more of its achieved task performance.

Table 5: **Additional results on instruction-tuned backbones.** We report Avg@32 (%) for the best checkpoint. Subtable (A) uses Qwen2.5-7B-Instruct, Subtable (B) uses Qwen2.5-Math-7B-Instruct, and Subtable (C) uses Qwen2.5-VL-7B-Instruct.

(A) Qwen2.5-7B-Instruct

Method	MATH500	HMMT25	AMC23	AIME25	AIME24	Avg.
Baseline	70.0 (+0.0)	5.6 (+0.0)	52.4 (+0.0)	3.5 (+0.0)	16.4 (+0.0)	29.6 (+0.0)
GRPO	79.2 (+9.2)	7.8 (+2.2)	68.3 (+15.9)	8.6 (+5.1)	27.4 (+11.0)	38.3 (+8.7)
GSPO	80.6 (+10.6)	8.5 (+2.9)	70.1 (+17.7)	10.8 (+7.3)	26.8 (+10.4)	39.4 (+9.8)
DAPO	81.2 (+11.2)	9.1 (+3.5)	71.3 (+18.9)	11.4 (+7.9)	28.0 (+11.6)	40.2 (+10.6)
SAPO	80.4 (+10.4)	8.0 (+2.4)	69.2 (+16.8)	10.1 (+6.6)	27.1 (+10.7)	39.0 (+9.4)
MHPO (Ours)	**82.1** (+12.1)	**13.8** (+8.2)	**74.6** (+22.2)	**16.7** (+13.2)	**31.5** (+15.1)	**43.7** (+14.1)

(B) Qwen2.5-Math-7B-Instruct

Method	MATH500	HMMT25	AMC23	AIME25	AIME24	Avg.
Baseline	73.1 (+0.0)	7.5 (+0.0)	56.3 (+0.0)	7.7 (+0.0)	15.1 (+0.0)	31.9 (+0.0)
GRPO	82.3 (+9.2)	9.6 (+2.1)	72.4 (+16.1)	11.0 (+3.3)	30.8 (+15.7)	41.2 (+9.3)
GSPO	83.8 (+10.7)	10.9 (+3.4)	74.6 (+18.3)	14.2 (+6.5)	30.1 (+15.0)	42.7 (+10.8)
DAPO	84.4 (+11.3)	11.5 (+4.0)	75.8 (+19.5)	15.0 (+7.3)	31.2 (+16.1)	43.6 (+11.7)
SAPO	83.6 (+10.5)	10.0 (+2.5)	72.5 (+16.2)	13.1 (+5.4)	30.0 (+14.9)	41.8 (+9.9)
MHPO (Ours)	**85.2** (+12.1)	**17.9** (+10.4)	**79.3** (+23.0)	**21.6** (+13.9)	**35.1** (+20.0)	**47.8** (+15.9)

(C) Qwen2.5-VL-7B-Instruct

Method	MathVista	MathVision	MathVerse	Avg.
Baseline	67.8 (+0.0)	25.2 (+0.0)	49.5 (+0.0)	47.5 (+0.0)
GRPO	68.6 (+0.8)	27.3 (+2.1)	52.7 (+3.2)	49.5 (+2.0)
GSPO	69.1 (+1.3)	27.4 (+2.2)	52.1 (+2.6)	49.5 (+2.0)
DAPO	71.2 (+3.4)	29.1 (+3.9)	56.5 (+7.0)	52.3 (+4.8)
SAPO	70.7 (+2.9)	28.6 (+3.4)	56.3 (+6.8)	51.9 (+4.4)
MHPO (Ours)	**71.8** (+4.0)	**30.5** (+5.3)	**56.7** (+7.2)	**53.0** (+5.5)

Table 6: **Training stability comparison between the best and latest checkpoints on Qwen3-4B-Base.** We report Avg@32 (%) with separate columns for the best checkpoint and the final checkpoint. Δ denotes the change from best to latest. MHPO shows the smallest degradation.

Method	MATH500		HMMT25		AMC23		AIME25		AIME24		Avg.		Δ
	Best	Latest	Best	Latest	Best	Latest	Best	Latest	Best	Latest	Best	Latest	
GRPO	64.8	58.9	11.1	8.8	81.3	72.5	30.8	23.3	40.4	24.2	45.7	37.5	−8.2
GSPO	66.2	59.9	12.4	8.6	83.5	75.9	34.0	25.9	39.7	27.0	47.2	39.5	−7.7
DAPO	64.0	47.5	14.9	7.0	75.8	66.3	29.2	17.9	34.2	18.4	43.6	31.4	−12.2
SAPO	66.4	59.2	11.5	7.3	81.4	77.6	32.9	23.4	39.6	24.0	46.4	38.3	−8.1
CISPO	69.0	63.2	11.0	7.8	85.3	78.4	33.5	26.0	43.7	32.5	48.5	41.6	−6.9
MASPO	67.5	60.8	12.8	8.3	82.6	76.0	34.5	26.8	40.8	29.1	47.6	40.2	−7.4
MHPO (Ours)	**70.9**	**70.1**	**19.4**	**19.4**	**88.2**	**88.2**	**41.4**	**40.8**	**44.7**	**41.8**	**52.9**	**52.1**	**−0.8**

F LIMITATIONS

The empirical scope is verifiable mathematical reasoning on Qwen-derived text and vision-language backbones. Mechanism diagnostics use Qwen3-4B, and the reported prefix-collapse comparisons focus on correct responses. Four DHP parameters are shared across the five original backbones, while the LFM range c is backbone-dependent. The sensitivity analysis varies one factor at a time.

The gradient moment certificate is local to the current parameter value and requires the stated advantage–score product-moment condition. DHP attenuates direct contributions according to ratio direction, including corrective updates when $\hat{A}_t > 0$ and $z_t < 0$. Shared-parameter interactions and optimizer state determine how these contributions combine into realized probability changes, as detailed in Appendix C.2.