FEPO: Future-Entropy Policy Optimization with Selective Update Modulation

Anonymous ACL submission

Abstract

Entropy is a fundamental measure of uncertainty. Existing RL methods for language models typically use token-level entropy, which only captures uncertainty at the local position and not over future generation. We argue this local quantity cannot fully characterize uncertainty in autoregressive decoding, and therefore focus on *future entropy*, i.e., uncertainty over the remaining suffix conditioned on the current prefix. By the entropy chain rule, local entropy is the expected decrease in future entropy, while the realized decrease depends on the chosen token, yielding an action-dependent future-entropy gap. We incorporate this signal into inference via MIN-F-MC, an oracle-style probe that selects the candidate token with the lowest estimated future-entropy rate at high-entropy positions. On mathematical reasoning tasks, MIN-F-MC improves reasoning performance, indicating that future entropy tracks reward-relevant reasoning choices. Motivated by this, we propose **FEPO** (**F**uture **E**ntropy **P**olicy **O**ptimization), which uses realized future-entropy rates at a sparse set of high-entropy positions to scale update magnitudes. By intervening only at a few critical positions, FEPO substantially improves mathematical reasoning performance.[1]

1 Introduction

Reinforcement learning with verifiable rewards (RLVR) has become a key paradigm for improving language-model reasoning on mathematics, tool use, and planning (Ouyang et al., 2022; Nakano et al., 2021; Shao et al., 2024; DeepSeek-AI, 2025). Group-based methods such as GRPO are especially practical because they avoid a learned value head, but they still broadcast the same sequence-level advantage to every token. This coarse update treats routine tokens and consequential reasoning decisions alike, echoing long-standing sequence-level training issues (Ranzato et al., 2016; Bengio et al., 2015). Prior work has introduced finer-grained training signals from process verifiers (Lightman et al., 2023; Li and Li, 2024), preference-based token objectives (Zheng et al., 2023; Rafailov et al., 2023; Zhong et al., 2024), and Monte Carlo or segment-level estimators (Kazemnejad et al., 2024; Guo et al., 2025), but these often require extra supervision, auxiliary models, or additional rollouts.

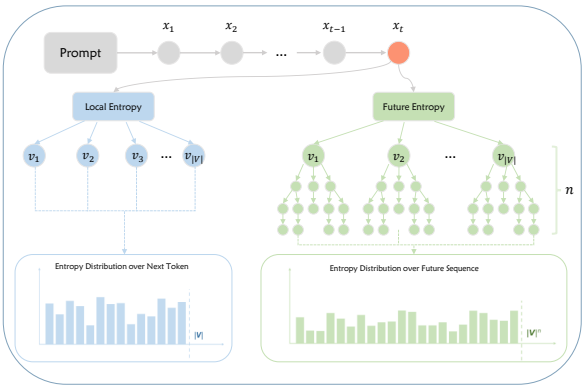

Figure 1: Future entropy vs. local entropy.

Recent evidence suggests a lower-cost path: update magnitude need not be uniform. High-entropy minority tokens can drive much of the learning signal (Wang et al., 2025a), related divergent-token patterns appear in efficient routing systems (Fu et al., 2025), and RL-induced reasoning gains concentrate in sparse critical positions (Meng et al., 2026). FEPO takes this sparse-intervention perspective while avoiding external process labels or counterfactual rollout trees.

Entropy is a natural uncertainty signal (Shannon, 1948; Cover, 1999), and recent RL methods use policy entropy for exploration, entropy control, or token-level update shaping (Tan et al., 2025; Xiong et al., 2026; Ahmed and Singh, 2026). Yet token entropy H_t (1) is local: it measures next-token uncertainty before an action is chosen, not the downstream uncertainty induced by that choice.

[1] Code is included in the Supplementary Material.

Consequently, selecting high-entropy positions using token-level entropy alone ignores how different choices at those positions reshape future uncertainty, a signal that we later show is behaviorally and training-time useful. We therefore introduce *future entropy* (5), the conditional entropy of the remaining suffix given the current prefix. The entropy chain rule shows that token entropy is only the expected future-entropy decrease, while the realized decrease also contains an action-dependent future-entropy gap $B_t(x_t)$ (7).

We first test this signal with MIN-F-MC (Figure 2, Section 3.3), an oracle-style inference probe that selects low-future-entropy candidates at high-entropy positions and improves over greedy decoding on math tasks. Motivated by this result, FEPO converts the same preference into a training-time rank relaxation: it reuses realized suffix-entropy rates from existing GRPO rollouts to modulate token-update *magnitude* on the low-entropy tail of sparse high-entropy positions, making future entropy a lightweight intervention signal for RLVR.

Our contributions are:

- We formulate *future entropy* for autoregressive reasoning and show that the realized future-entropy decrease decomposes into the familiar token entropy and an action-dependent future-entropy gap, clarifying what token-level entropy cannot express.
- We introduce MIN-F-MC, an oracle-style inference probe that uses Monte Carlo continuation sampling to test whether lower future entropy at high-entropy decision points is aligned with better mathematical reasoning outcomes.
- We propose FEPO, a training-time update modulator that ranks rollout-level suffix-entropy rates and adjusts the magnitude of GRPO advantages on the low-entropy tail of sparse high-entropy positions.

2 Related Work

2.1 GRPO and the credit-assignment gap

Group Relative Policy Optimization (GRPO) has become a practical recipe for RLVR because it estimates trajectory-level advantages by normalizing rewards across multiple completions sampled for the same prompt, thereby avoiding the learned value function used in PPO-style training (Schulman et al., 2017; Shao et al., 2024; DeepSeek-AI, 2025). Recent work further refines GRPO along several axes, including reward normalization, length effects, token aggregation, success amplification, and critic-free credit propagation (Yu et al., 2025; Liu et al., 2025; Mroueh, 2025; Parthasarathi et al., 2025; Wang et al., 2025b). However, the group-relative advantage is still a sequence-level signal that is broadcast to all tokens in the rollout. As a result, every token receives the same update direction and magnitude even though, in long-chain reasoning, different tokens can have very different importance and only a small subset of decisions may actually determine the final answer.

2.2 Entropy as a signal: from local to future uncertainty

Token entropy has been used in RLVR primarily in two roles. The first is policy-stability control: entropy regularization to avoid premature collapse (Williams and Peng, 1991; Mnih et al., 2016; Meister et al., 2020), and entropy-aware clipping mechanisms to manage the well-known entropy dynamics of PPO-style training (Yu et al., 2025; Cui et al., 2025; Lei and Baehr, 2026). The second is token-level update shaping: GTPO and its variants reweight reasoning steps by local token entropy, and GRPO-S and ETR use token-entropy-based reshaping of the trajectory reward to emphasize high-uncertainty branches (Tan et al., 2025; Cheng et al., 2025; Xiong et al., 2026; Cao et al., 2025). A common feature of all of these methods is that the entropy quantity used is *local*: it depends only on the next-token distribution at the current prefix.

EPIC (Ahmed and Singh, 2026) incorporates a finite-horizon Rao–Blackwellized lookahead future entropy into *decoding* and aligns the sampling distribution's stepwise uncertainty with the estimated downstream uncertainty. Test-time uncertainty- or entropy-control during decoding more broadly has also been studied in information-seeking and target-perplexity settings (Hu et al., 2024; Basu et al., 2020). By contrast, FEPO operates at training time and uses future-entropy information as a within-rollout magnitude modulator on top of the GRPO advantage; we make this distinction precise in Section 4.

3 Theory

In this section, we first define the GRPO setup, then introduce future entropy and its relationship to token-level entropy.

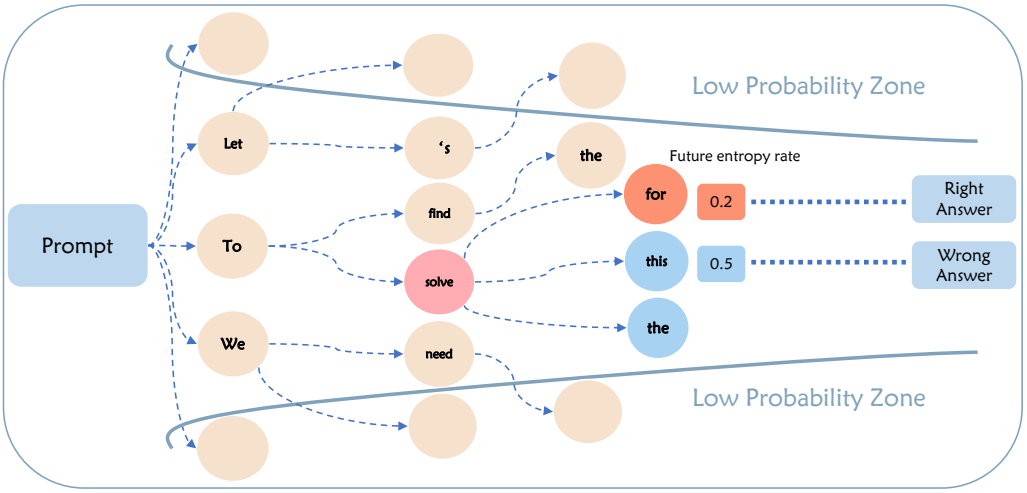

Figure 2: Overview of the MIN-F-MC inference probe. At high-entropy positions, the probe branches over top-k candidate tokens, estimates each candidate's downstream future entropy with Monte Carlo continuations, and selects the candidate with the lowest estimate. Other positions fall back to greedy decoding.

3.1 Setup and notation

Consider an autoregressive language model over vocabulary $\mathcal{V}$ that generates $x = (x_1, \ldots, x_n)$. Let $x_{<t} = (x_1, \ldots, x_{t-1})$ and $x_{\le t} = (x_1, \ldots, x_t)$. At position t, write $p_t(v) \triangleq p(v \mid x_{<t})$ for the next-token distribution and define the *token entropy*

$$\begin{aligned} H_t &\triangleq H(X_t \mid x_{<t}) \\ &= -\sum_{v \in \mathcal{V}} p_t(v) \log p_t(v). \end{aligned} \tag{1}$$

GRPO objective. For a prompt q, GRPO (Shao et al., 2024) samples K rollouts $\{x^{(k)}\}_{k=1}^{K}$ from π_{old} and assigns each a scalar reward $R^{(k)}$. The group-normalized trajectory advantage is

$$\begin{aligned} \bar{A}^{(k)} &\triangleq \frac{R^{(k)} - \mu_R}{\sigma_R + \varepsilon}, \\ \mu_R &= \tfrac{1}{K} \sum_k R^{(k)}, \\ \sigma_R^2 &= \tfrac{1}{K} \sum_k \big(R^{(k)} - \mu_R\big)^2, \end{aligned} \tag{2}$$

which is broadcast to every token: $A_t^{(k)} = \bar{A}^{(k)}$. With ratio $r_t^{(k)}(\theta) = \pi_\theta(x_t^{(k)} \mid x_{<t}^{(k)}, q) / \pi_{\text{old}}(x_t^{(k)} \mid x_{<t}^{(k)}, q)$ and clipped token objective

$$\begin{aligned} \ell_t^{(k)}(\theta) \triangleq \min\Big\{ & r_t^{(k)}(\theta)\, A_t^{(k)}, \\ & \text{clip}\big(r_t^{(k)}(\theta),\, 1-\epsilon,\, 1+\epsilon\big)\, A_t^{(k)} \Big\}, \end{aligned} \tag{3}$$

the GRPO loss is

$$\begin{aligned} \mathcal{L}_{\text{GRPO}}(\theta) = -\,\mathbb{E}_{q,\{x^{(k)}\}} \Bigg[& \frac{1}{K} \sum_{k=1}^{K} \frac{1}{n^{(k)}} \\ & \cdot \sum_{t=1}^{n^{(k)}} \ell_t^{(k)}(\theta) \Bigg] \\ & + \beta_{\text{KL}}\, \mathcal{D}_{\text{KL}}\big(\pi_\theta \,\|\, \pi_{\text{ref}}\big), \end{aligned} \tag{4}$$

with the standard PPO clip and optional KL regularization toward π_{ref}. FEPO modifies only $A_t^{(k)}$ in (3); the importance ratio, the clip, and the KL term are unchanged.

3.2 Future entropy and the bias decomposition

Definition 1 (Future entropy)**.** For any prefix s, the *future entropy* is the joint conditional entropy of the remaining suffix,

$$\begin{aligned} F(s) &\triangleq H\big(X_{|s|+1}, \ldots, X_n \mid s\big), \\ F(x_{\le n}) &= 0, \end{aligned} \tag{5}$$

and the *averaged future entropy* after position t is

$$\bar{F}_t \triangleq \mathbb{E}_{V \sim p_t}\big[F(x_{<t}, V)\big]. \tag{6}$$

Unlike H_t, $F(s)$ measures uncertainty over the whole remaining trajectory.

The chain rule gives:

Proposition 1 (Bias decomposition of the per-step entropy change)**.** *Let* $\delta_t(x_t) \triangleq F(x_{<t}) - F(x_{\le t})$

and $B_t(x_t) \triangleq \bar{F}_t - F(x_{\le t})$. *Then*

$$\begin{aligned} F(x_{<t}) &= H_t + \bar{F}_t, \\ \delta_t(x_t) &= H_t + B_t(x_t), \\ \mathbb{E}_{X_t \sim p_t}[B_t(X_t)] &= 0. \end{aligned} \tag{7}$$

Thus H_t *is the expected future-entropy reduction, while* B_t *is the action-dependent residual averaged out by* H_t*. The proofs are given in Appendices A.1–A.4.*

Proposition 2 (Unbiased suffix-entropy estimator)**.** *For prefix* $x_{\le t}$*, sample one suffix from* $p(\cdot \mid x_{\le t})$ *and compute* $H_s \triangleq H(X_s \mid x_{<s})$ *on the realized path. Then*

$$\begin{aligned} \widehat{F}_t &\triangleq \sum_{s=t+1}^{n} H_s, \\ \widehat{E}_t &\triangleq \widehat{F}_t/(n-t). \end{aligned} \tag{8}$$

Under on-policy sampling, $\mathbb{E}[\widehat{F}_t \mid x_{\le t}] = F(x_{\le t})$ *and* $\mathbb{E}[\widehat{E}_t \mid x_{\le t}] = F(x_{\le t})/(n-t)$*; see Appendix A.5.*

The bias term B_t is the only action-dependent component of $\delta_t(x_t)$. Estimating its ordering across candidate tokens requires same-prefix counterfactual continuations, making direct online recovery expensive in long-horizon RL.

3.3 Test-time probe: MIN-F-MC

To test whether lower future entropy is reward-aligned, MIN-F-MC intervenes only at positions with $H_t > \tau$. It forms $\mathcal{C}_t$ from the top-k tokens of p_t, samples M continuations for each $v \in \mathcal{C}_t$, and estimates

$$\widehat{F}_{\mathrm{MC}}(x_{<t}, v) = \frac{1}{M} \sum_{m=1}^{M} \frac{1}{\min(H_{\max},\, n_m - t)} \cdot \sum_{s=t+1}^{t+\min(H_{\max},\, n_m - t)} H_s^{(m)}, \tag{9}$$

This length-normalized rate is used to select $x_t^\star = \arg\min_{v \in \mathcal{C}_t} \widehat{F}_{\mathrm{MC}}(x_{<t}, v)$; other positions fall back to greedy. For a fixed prefix, this selects the candidate with the largest empirical future-entropy gap among the candidates.

Random TopK uses the same routed positions but samples from the renormalized top-k distribution, isolating the future-entropy criterion from the effect of intervening itself. Across both Qwen3-4B and Qwen3-8B, and on both Math 500 and the training-distribution subset D, MIN-F-MC improves over greedy and Random TopK (Table 1), supporting a training-time rank relaxation that favors low realized suffix entropy in existing rollouts. The 4B/8B absolute ordering is not monotonic here, so we interpret the probe by within-checkpoint gains over greedy.

4 FEPO

Section 3.3 shows how future entropy can guide sparse interventions at inference time through MIN-F-MC. We now translate the same preference to RL training: FEPO reuses realized suffix-entropy rates from existing GRPO rollouts to modulate update magnitude at sparse high-entropy positions while avoiding external process labels and counterfactual rollout trees.

4.1 Working assumptions

Assumption 1 (Sparse vital-few)**.** Reward-relevant decisions in RLVR concentrate in a small high-token-entropy set $\mathcal{T} = \{(k,t) : H_t^{(k)} > \tau\}$, where τ is an entropy threshold; conditioning on $\mathcal{T}$ captures positions where selective update modulation is most likely to matter (Meng et al., 2026).

Assumption 2 (Low-future-entropy / large-gap alignment)**.** Within $\mathcal{T}$, samples with lower realized future entropy, and hence larger empirical future-entropy gaps among comparable samples, are more reward-aligned on average. Operationally, $\mathbb{P}(A_t^{(k)} > 0 \mid q_t^{(k)})$ is higher in the low-q tail than in the high-q head of the batch-level quantile $q_t^{(k)}$ of the length-debiased score (Definition 2).

A1 is supported by recent RLVR diagnostics; A2 is probed by MIN-F-MC and its in-training analogue (Figure 4). Together, these assumptions make the FEPO modulator well-defined and stable.

Off-policy suffix scoring. Proposition 2 is exact only when the suffix is sampled and scored under the same policy. In FEPO, rollouts are drawn from π_{old} whereas suffix entropies are scored under the current policy π_θ. Thus the future-entropy estimate in Equation (10) is biased under off-policy sampling. However, we show and empirically verify that this bias is controllable for FEPO's rank-based stability; Assumption 3 (A3) and Proposition 3 make this statement precise.

4.2 Training-time update modulator: FEPO

Length-bucket normalization. FEPO ranks realized suffix-entropy rates across high-entropy po-

Table 1: Test-time probe comparison on MATH 500 and D (a 300-example subset drawn from the training set). In MIN-F-MC, future-entropy rate is estimated with 128 Monte Carlo continuations per candidate token. Random TopK samples from the top-k candidate set at high-entropy positions after renormalizing their probabilities. Δ is computed against greedy under the same model and dataset.

Model	**Method**	**Math 500**				**D**			
		Acc.	Δ	**Branch**	**Length**	**Acc.**	Δ	**Branch**	**Length**
Qwen3-4B	Greedy	42.0	+0.0	0	721.7	19.0	+0.0	0	1036.7
	Random TopK	43.6	+1.6	39.5	732.7	17.6	-1.4	25.9	1076.2
	MIN-F-MC	51.6	+9.6	18.1	860.6	30.0	+11.0	24.5	1200.2
Qwen3-8B	Greedy	40.6	+0.0	0	675.8	29.6	+0.0	0	1460.5
	Random TopK	31.6	-9.0	30.4	709.8	20.6	-9.0	25.3	1385.6
	MIN-F-MC	47.4	+6.8	11.7	820.4	35.3	+5.7	47.9	1609.8

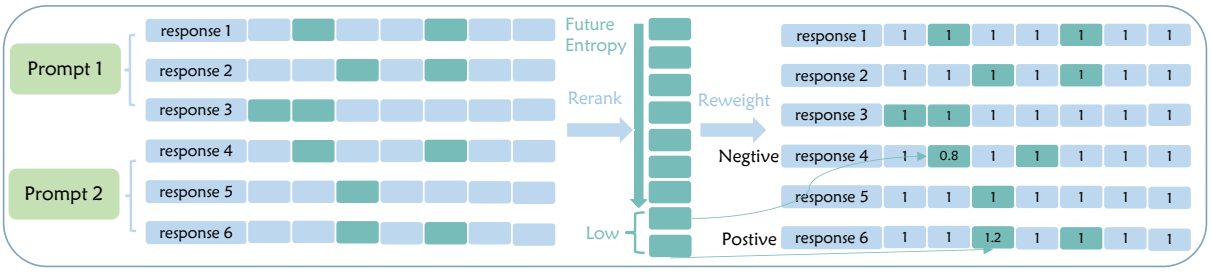

Figure 3: Overview of FEPO. At sparse high-entropy positions, suffix entropy is estimated from rollouts and used to modulate token-level advantage magnitude via a low-tail rank-based selective update signal.

sitions in a batch. Because remaining suffix lengths vary across positions and rollouts, we z-score each rate within a suffix-length bucket before ranking.

Definition 2 (Suffix-entropy signal and length-debiased score)**.** For rollout k and position t, define

$$\widehat{E}_t^{(k)} \triangleq \frac{1}{n^{(k)} - t} \sum_{s=t+1}^{n^{(k)}} H_s^{(k)}, \qquad Z_t^{(k)} \triangleq \frac{\widehat{E}_t^{(k)} - \mu_{b(L_t^{(k)})}}{\sigma_{b(L_t^{(k)})} + \varepsilon}, \tag{10}$$

where $L_t^{(k)} = n^{(k)} - t$, $b(\cdot)$ maps suffix lengths to fixed-width buckets, and (μ_b, σ_b^2) are running bucket statistics (Appendix C). In our implementation, each bucket covers three consecutive suffix lengths. Let $q_t^{(k)} \in [0, 1]$ be the percentile rank of $Z_t^{(k)}$ within the high-entropy set $\mathcal{T}$ after applying the per-rollout cap ρ_{cap}. Lower $q_t^{(k)}$ indicates lower length-debiased suffix entropy, i.e., a larger empirical future-entropy gap among the ranked training samples.

Definition 3 (FEPO modulator)**.** Given the GRPO advantage $A_t^{(k)}$, FEPO applies a sign-asymmetric multiplier on the low-q tail:

$$m_t^{(k)} = \begin{cases} 1 + \alpha_{\text{pos}}, & q_t^{(k)} \le \beta,\ A_t^{(k)} > 0, \\ 1 - \alpha_{\text{neg}}, & q_t^{(k)} \le \beta,\ A_t^{(k)} < 0, \\ 1, & \text{otherwise}, \end{cases} \tag{11}$$

$$\widetilde{A}_t^{(k)} \triangleq m_t^{(k)}\, A_t^{(k)},$$

where $\alpha_{\text{pos}} \ge 0$, $\alpha_{\text{neg}} \ge 0$, and $\beta \in (0, 1/2)$ is a fixed constant. $\widetilde{A}_t^{(k)}$ replaces $A_t^{(k)}$ inside the GRPO surrogate (3); the importance ratio, the clip, and the KL term are unchanged.

Why this asymmetry. Under Assumption 2 (A2), low-q positions on positive-advantage rollouts tend to carry larger empirical future-entropy gaps and be reward-aligned, so FEPO amplifies their gradients. On negative-advantage rollouts, such positions may be locally valid steps inside an ultimately wrong trajectory; FEPO therefore attenuates rather than amplifies their negative gradients (Section 5.4). The high-suffix-entropy head is left unchanged because suppressing it removes useful exploratory branches (Section 5.4).

Proposition 3 (Low-tail stability under bounded off-policy shift)**.** *Under Assumption 3 (A3), global proximity between π_{old} and π_θ controls the expected score drift used by FEPO. Consequently,*

the off-policy low-tail set inside $\mathcal{T}$ can differ from the oracle current-policy low tail only at a Markov-controlled set of large-drift positions or near the empirical β-quantile boundary.

Proof in Appendix B. Combined with Assumption 1, FEPO turns $\mathcal{T}$ into a stable rank-based selective update signal, inherits its update direction from GRPO, and adds only a backward cumulative sum plus lightweight bucket statistics—no extra forward passes, Monte Carlo branching, or auxiliary models. At training time, same-prefix counterfactual branching is replaced by a batch-level rank relaxation: low-q positions are samples with the largest empirical future-entropy gaps available from existing rollouts. Table 3 quantifies the low-q/large-B_t alignment by comparing increasingly narrow low-q cuts with an expensive MC=128 same-prefix gap-direction reference; the agreement rises as q becomes smaller and at the default $\beta = 0.2$ is comparable to a single-sample MC estimator. FEPO remains robust because the signal is sparse, the modulation magnitude is moderate, and the GRPO advantage retains control of the update sign.

5 Experiments

5.1 Experimental Setup

Tasks and Datasets. We train on dapo17k (Yu et al., 2025) and evaluate on AIME24 and AIME25 (Mathematical Association of America, 2026a), MATH 500 (Lightman et al., 2023), AMC (Mathematical Association of America, 2026b), OlympiadBench (He et al., 2024), and Minerva (Lewkowycz et al., 2022). **Evaluation Metrics.** Our primary evaluation metrics are Pass@32 and Mean@32. **Models and Baselines.** We evaluate Qwen3-4B and Qwen3-8B from the Qwen3 family (Yang et al., 2025), and compare FEPO against GRPO, GTPO, GRPO-S and ETR. For all baselines, method-specific hyperparameters are set to the values reported as optimal in the respective source papers, so that comparisons reflect each method under its recommended configuration. **Implementation details.** Training hyperparameters, optimizer settings, and generation settings are reported in Appendix E. **Artifact documentation.** Artifact sources, use conditions, intended use, and data statistics are summarized in Appendix F. **FEPO defaults.** Unless otherwise stated, FEPO uses $\alpha_{\text{pos}} = \alpha_{\text{neg}} = 0.2$, tail fraction $\beta = 0.2$, entropy threshold $\tau = 1.0$, and per-response cap $\rho_{\text{cap}} = 5\%$ of the response length. For length debiasing, FEPO uses bucket-wise z-scoring with fixed-width suffix-length buckets $b(L) = \lfloor (L-1)/3 \rfloor$ for $L \geq 1$, minimum bucket count $N_{\min} = 5$, EMA rate $\eta = 0.05$, and score clipping $c = 3$.

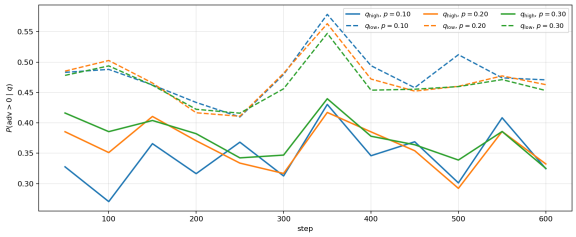

Figure 4: Suffix-entropy quantile diagnostic. We compare $\mathbb{P}(A_t^{(k)} > 0 \mid q_t^{(k)})$ across low-tail and high-head suffix-entropy quantiles in the high-entropy set $\mathcal{T}$.

5.2 Main results

We compare FEPO against four representative GRPO-style and entropy-aware update baselines: GRPO (group-relative advantage), GTPO (token-entropy-weighted updates), GRPO-S (entropy-shaped sequence-level updates), and ETR (entropy-augmented terminal reward); Mean@32 (M@32) and Pass@32 (P@32) on six math benchmarks are reported in Table 2, with AIME evaluation curves in Figure 5.

FEPO is the strongest overall method, leading on 11/12 Mean@32 cells and improving over GRPO across all benchmarks and both model sizes. Its Pass@32 gains are most visible on AIME24/AIME25, where selective update modulation improves hard-reasoning success rates. The AIME curves in Figure 5 further show faster convergence without noticeably noisier training. The Base rows show the same non-monotonic checkpoint behavior, so we treat Qwen3-4B and Qwen3-8B as related but distinct backbones and emphasize within-backbone comparisons.

5.3 Diagnostics

(A) Realized analogue of MIN-F-MC inside $\mathcal{T}$. Figure 4 shows lower realized future entropy correlates with higher positive-advantage probability inside $\mathcal{T}$, supporting Assumption 2 (A2) and the choice to amplify the low tail.

Table 3 reports how strongly the FEPO low-q proxy agrees with the MC=128 same-prefix large-B_t direction. The trend is monotonic in the intended direction: agreement is highest for the tightest low-q cut (0.81 at $q \leq 0.05$) and decreases as

Table 2: Main experiment results. Each benchmark reports Mean@32 (M@32) and Pass@32 (P@32); Average reports the mean over the six benchmarks separately for M@32 and P@32. Within each column and each model block, the best result among trained methods (GRPO, GTPO, GRPO-S, ETR, FEPO) is shown in **bold** and the second best is underlined

Method	AIME24		AIME25		MATH 500		AMC		Olympiad		Minerva		Average	
	M@32	P@32	M@32	P@32	M@32	P@32	M@32	P@32	M@32	P@32	M@32	P@32	M@32	P@32
Qwen3-4B														
Base	6.1	36.7	5.3	33.3	39.7	90.8	26.4	82.5	21.7	66.5	12.9	44.5	18.7	59.1
GRPO	16.4	43.3	11.5	40.0	79.8	93.0	58.9	92.5	44.1	68.9	40.6	61.4	41.9	66.5
GTPO	16.5	36.7	13.5	38.0	74.8	89.4	58.6	87.5	39.5	61.9	37.4	62.6	40.1	62.7
GRPO-S	18.8	43.3	16.7	45.6	83.5	**96.2**	**64.8**	92.5	43.2	64.4	40.9	59.2	44.6	66.9
ETR	16.4	36.7	12.1	40.0	79.7	95.0	57.9	**95.0**	46.4	**71.1**	38.6	57.4	41.9	65.9
FEPO	**19.8**	**50.0**	**17.8**	**46.7**	**85.6**	**96.2**	63.5	90.0	**49.6**	68.0	**42.8**	**64.7**	**46.5**	**69.3**
Qwen3-8B														
Base	5.8	40	4.9	30	40.8	92	24.6	92.5	18.9	67.4	10.9	49.3	17.7	61.9
GRPO	25.3	60.0	17.9	56.7	86.6	97.2	67.0	95.0	54.1	76.1	43.6	62.1	49.1	74.5
GTPO	26.4	**73.3**	16.7	50	82.7	93.2	68.6	**97.5**	50.1	71.2	41.7	63.3	47.7	74.8
GRPO-S	27.6	60.0	18.9	46.7	85.8	97.2	69.1	95.0	49.6	74.6	41.0	**66.7**	48.7	73.4
ETR	26.7	66.7	15.6	46.7	84.1	97.2	64.0	92.5	51.8	77.1	41.1	63.2	47.2	73.9
FEPO	**30.1**	**73.3**	**20.3**	**66.7**	**87.6**	**97.4**	**72.0**	95.0	**56.0**	**79.2**	**44.3**	65.6	**51.7**	**79.5**

Table 3: Low-q alignment with a large-B_t reference. MC=128 same-prefix continuations estimate the future-entropy gap direction; for increasingly narrow FEPO low-q selections, we report agreement with this reference. Higher agreement at smaller q indicates that the low tail increasingly concentrates samples whose direction matches larger B_t. Lookahead and MC=1 are included as alternative low-cost references.

Estimator	Selection	Large-B_t agree.
FEPO proxy	$q \leq 0.05$	0.81
	$q \leq 0.10$	0.76
	$q \leq 0.20$	0.72
	$q \leq 0.30$	0.67
	$q \leq 0.50$	0.56
Lookahead	1 step	0.64
Lookahead	2 steps	0.65
Monte Carlo	MC=1	0.73

the tail widens. At the default $\beta = 0.2$, FEPO remains on par with MC=1 (0.72 vs. 0.73) and above short lookahead, supporting the use of low-q as a practical large-gap training signal.

(B) Off-policy: low-tail rank stability. Suffix entropies are scored under π_θ on π_{old} rollouts. This diagnostic isolates whether rescoring the same old-policy suffixes under π_θ preserves the ranking used by FEPO; fresh π_θ rollouts would measure the full new-policy future-entropy proxy. Appendix B first bounds the global average expected score shift from the average suffix KL $D_{\text{KL}}(\pi_{\text{old}} \| \pi_\theta)$ and then shows that the inherited large-drift positions plus an expanded empirical quantile-boundary band are the only places where low-tail membership can change inside $\mathcal{T}$. Figure 6 tests the rank-level consequence directly: it shows consistently high Spearman correlation between π_θ- and π_{old}-based scores on $\mathcal{T}$, supporting the stability conclusion in practice.

5.4 Ablations

We ablate the design choices in Section 4.2. Table 6 (Appendix G.1) reports each setting relative to the FEPO default. Across all sweeps, the default configuration is best. **Sign asymmetry (α_{neg}).** With α_{pos} fixed, attenuating negative-advantage gradients on the low tail ($\alpha_{\text{neg}} = 0.2$) beats one-sided positive-only modulation ($\alpha_{\text{neg}} = 0$) and negative amplification ($\alpha_{\text{neg}} < 0$), consistent with our goal of encouraging low-future-entropy choices with relatively large empirical future-entropy gaps. **High-head suppression.** Suppressing the high-suffix-entropy head ($\alpha_{\text{high}} \neq 0$) hurts in both directions, supporting the choice to leave the head unmodified: high suffix entropy can indicate either unconfident chaotic continuation or useful exploratory branching, so suppressing the high tail risks penalizing productive exploration. **Length-bucket correction.** Ranking by raw $\widehat{E}_t^{(k)}$ instead of $Z_t^{(k)}$ reduces accuracy, confirming that length would otherwise dominate batch-level ranking. **Tail fraction, gating, and cap (β, τ, ρ_{cap}).** Performance is stable for intermediate β (best at 0.2) but degrades when β is too small or too large; tighter entropy thresholds ($\tau = 0.8$ or top-10%) also hurt, and an intermedi-

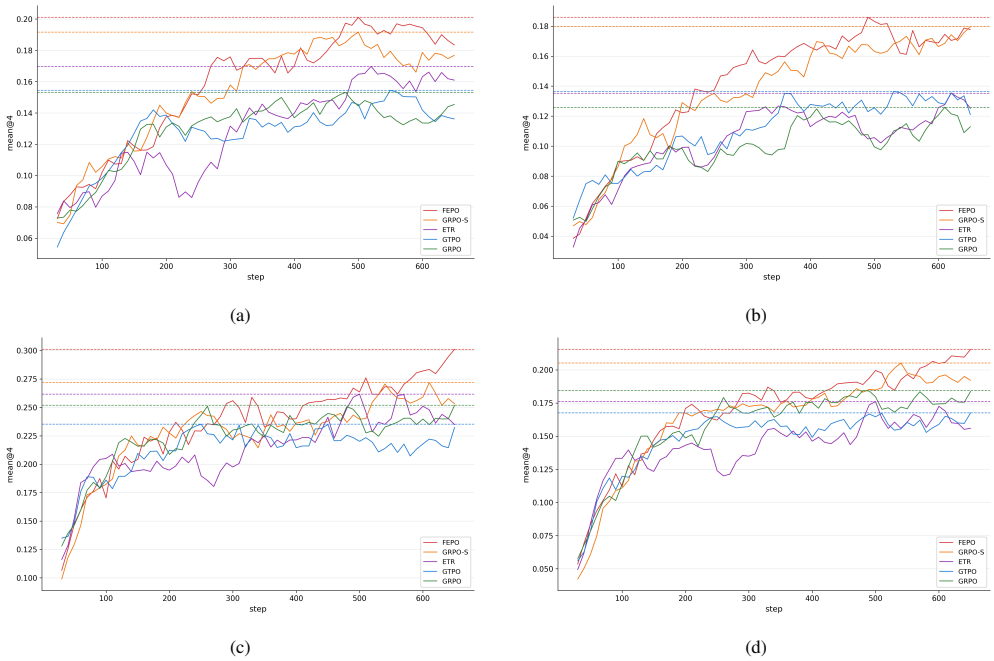

Figure 5: Evaluation curves on the AIME suite. Panels (a) and (b) show Qwen3-4B Mean@4 on AIME24 and AIME25; panels (c) and (d) show the corresponding Qwen3-8B Mean@4 curves.

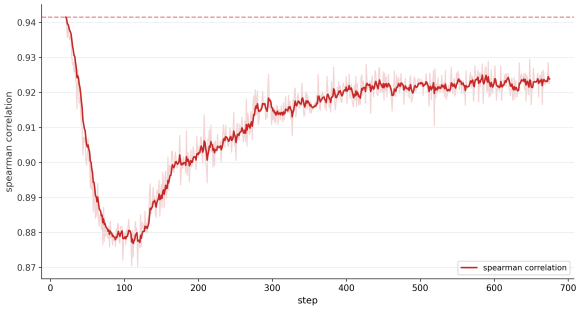

Figure 6: Off-policy diagnostic. Spearman rank correlation between π_θ-based and π_{old}-based suffix-entropy scores, averaged across prompts during training.

ate per-response cap ($\rho_{\text{cap}} = 5\%$) beats both no cap and overly tight caps.

6 Conclusion

This paper enriches our understanding of entropy from the perspective of *future entropy*. Token-level (local) entropy characterizes local uncertainty, whereas future entropy characterizes uncertainty over the longer-horizon future of the generation. At sparse high-entropy decision points, this longer-horizon signal further distinguishes among candidate continuations. Our MIN-F-MC probe shows that, at high-entropy decision points, selecting the candidate token with a relatively more certain future is more likely to lead to the correct answer. FEPO turns this same-prefix preference into a batch-level training relaxation: it uses realized suffix-entropy ranks as a lightweight proxy for relatively large empirical future-entropy gaps and modulates the GRPO advantage only on the low-suffix-entropy tail of the sparse high-entropy set, amplifying positive advantages and attenuating negative ones. The takeaway is that future entropy is a model-intrinsic signal that goes beyond local entropy: it reshapes *how* GRPO's reward gradients are weighted at the few high-entropy decision points that actually matter.

7 Limitations

Our study is currently validated only on mathematical reasoning tasks and on Qwen3-4B/8B backbones; transfer to code, tool use, planning, multilingual reasoning, and other model families remains to be tested. FEPO also relies on the empirical assumption that, within sparse high-token-entropy positions, lower realized suffix-entropy rates are more likely to align with positive reward advantages. Our diagnostics and ablations support this assumption in the studied setting, but it may not hold uniformly across tasks where exploration, diversity, or multiple valid reasoning paths are more important. During training, FEPO needs to estimate suffix-entropy rates for each selected high-entropy position, which introduces additional runtime overhead.

References

K. Ahmed and S. Singh. 2026. Entropy-aligned decoding of LMs for better writing and reasoning. *arXiv preprint arXiv:2601.01714*.

Sourya Basu, Govardana Sachitanandam Ramachandran, Nitish Shirish Keskar, and Lav R. Varshney. 2020. Mirostat: A neural text decoding algorithm that directly controls perplexity. *arXiv preprint arXiv:2007.14966*.

S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, volume 28, pages 1171–1179.

Hongye Cao, Zhixin Bai, Ziyue Peng, Boyan Wang, Tianpei Yang, Jing Huo, Yuyao Zhang, and Yang Gao. 2025. Efficient reinforcement learning with semantic and token entropy for LLM reasoning. *arXiv preprint arXiv:2512.04359*.

D. Cheng, S. Huang, X. Zhu, B. Dai, X. Zhao, Z. Zhang, and F. Wei. 2025. Reasoning with exploration: An entropy perspective. *arXiv preprint arXiv:2506.14758*.

Thomas M Cover. 1999. *Elements of information theory*. John Wiley & Sons.

G. Cui, Y. Zhang, J. Chen, L. Yuan, Z. Wang, Y. Zuo, H. Li, Y. Fan, H. Chen, W. Chen, Z. Liu, H. Peng, L. Bai, W. Ouyang, Y. Cheng, B. Zhou, and N. Ding. 2025. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

T. Fu, Y. Ge, Y. You, E. Liu, Z. Yuan, G. Dai, S. Yan, H. Yang, and Y. Wang. 2025. R2R: Efficiently navigating divergent reasoning paths with small-large model token routing. *arXiv preprint arXiv:2505.21600*.

Yiran Guo, Lijie Xu, Jie Liu, Dan Ye, and Shuang Qiu. 2025. Segment policy optimization: Effective segment-level credit assignment in RL for large language models. *arXiv preprint arXiv:2505.23564*.

C. He, R. Luo, Y. Bai, S. Hu, Z. L. Thai, J. Shen, J. Hu, X. Han, Y. Huang, Y. Zhang, J. Liu, L. Qi, Z. Liu, and M. Sun. 2024. OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*.

Z. Hu, C. Liu, X. Feng, Y. Zhao, S.-K. Ng, A. T. Luu, J. He, P. W. Koh, and B. Hooi. 2024. Uncertainty of thoughts: Uncertainty-aware planning enhances information seeking in large language models. *arXiv preprint arXiv:2402.03271*.

A. Kazemnejad, M. Aghajohari, E. Portelance, A. Sordoni, S. Reddy, A. Courville, and N. Le Roux. 2024. VinePPO: Refining credit assignment in RL training of LLMs. *arXiv preprint arXiv:2410.01679*.

Ming Lei and Christophe Baehr. 2026. A comparative theoretical analysis of entropy control methods in reinforcement learning for reasoning language models. *arXiv preprint arXiv:2604.09676*.

A. Lewkowycz, A. Andreassen, D. Dohan, E. Dyer, H. Michalewski, V. Ramasesh, A. Slone, C. Anil, I. Schlag, T. Gutman-Solo, Y. Wu, B. Neyshabur, G. Gur-Ari, and V. Misra. 2022. Solving quantitative reasoning problems with language models. *arXiv preprint arXiv:2206.14858*.

Wendi Li and Yixuan Li. 2024. Process reward model with Q-value rankings. *arXiv preprint arXiv:2410.11287*.

H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever, and K. Cobbe. 2023. Let's verify step by step. *arXiv preprint arXiv:2305.20050*.

Z. Liu, C. Chen, W. Li, P. Qi, T. Pang, C. Du, W. S. Lee, and M. Lin. 2025. Understanding R1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*.

Mathematical Association of America. 2026a. AIME: American invitational mathematics examination. https://maa.org/maa-invitational-competitions. Accessed 2026.

Mathematical Association of America. 2026b. AMC 10/12: American mathematics competitions. https://maa.org/maa-invitational-competitions. Accessed 2026.

C. Meister, R. Cotterell, and T. Vieira. 2020. If beam search is the answer, what was the question? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2173–2185.

Haoming Meng, Kexin Huang, Shaohang Wei, Chiyu Ma, Shuo Yang, Xue Wang, Guoyin Wang, Bolin Ding, and Jingren Zhou. 2026. Sparse but critical: A token-level analysis of distributional shifts in RLVR fine-tuning of LLMs. *arXiv preprint arXiv:2603.22446*.

V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu. 2016. Asynchronous methods for deep reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 1928–1937.

Y. Mroueh. 2025. Reinforcement learning with verifiable rewards: GRPO's effective loss, dynamics, and success amplification. *arXiv preprint arXiv:2503.06639*.

R. Nakano, J. Hilton, S. Balaji, J. Wu, L. Ouyang, C. Kim, C. Hesse, S. Jain, V. Kosaraju, W. Saunders, X. Jiang, K. Cobbe, T. Eloundou, G. Krueger, K. Button, M. Knight, B. Chess, and J. Schulman. 2021. WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

P. Parthasarathi, M. Reymond, B. Chen, Y. Cui, and S. Chandar. 2025. GRPO-λ: Credit assignment improves LLM reasoning. *arXiv preprint arXiv:2510.00194*.

R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36.

M.-A. Ranzato, S. Chopra, M. Auli, and W. Zaremba. 2016. Sequence level training with recurrent neural networks. In *International Conference on Learning Representations*.

J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

C. E. Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27(3–4):379–423, 623–656.

Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. K. Li, Y. Wu, and D. Guo. 2024. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

H. Tan, Z. Wang, J. Pan, J. Lin, H. Wang, Y. Wu, T. Chen, Z. Zheng, Z. Tang, and H. Yang. 2025. GTPO and GRPO-S: Token and sequence-level reward shaping with policy entropy. *arXiv preprint arXiv:2508.04349*.

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

Y. Wang, J. Zhao, C. Zhao, S. Guan, G. Penn, and S. Liu. 2025b. λ-GRPO: Unifying the GRPO frameworks with learnable token preferences. *arXiv preprint arXiv:2510.06870*.

Ronald J. Williams and Jing Peng. 1991. Function optimization using connectionist reinforcement learning algorithms. *Connection Science*, 3(3):241–268.

Xuan Xiong, Huan Liu, Li Gu, Zhixiang Chi, Yue Qiu, Yuanhao Yu, and Yang Wang. 2026. ETR: Entropy trend reward for efficient chain-of-thought reasoning. *arXiv preprint arXiv:2604.05355*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, and 16 others. 2025. DAPO: An open-source LLM reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.

Rui Zheng, Shihan Dou, Songyang Gao, Yuan Hua, Wei Shen, Binghai Wang, Yan Liu, Senjie Jin, Qin Liu, Yuhao Zhou, Limao Xiong, Lu Chen, Zhiheng Xi, Nuo Xu, Wenbin Lai, Minghao Zhu, Cheng Chang, Zhangyue Yin, Rongxiang Weng, and 8 others. 2023. Secrets of RLHF in large language models part I: PPO. *arXiv preprint arXiv:2307.04964*.

Han Zhong, Zikang Shan, Guhao Feng, Wei Xiong, Xinle Cheng, Li Zhao, Di He, Jiang Bian, and Liwei Wang. 2024. DPO meets PPO: Reinforced token optimization for RLHF. *arXiv preprint arXiv:2404.18922*.

A Proofs

This appendix contains the derivations used in the main text.

A.1 Proof of Proposition 1 (chain-rule identity)

By the entropy chain rule (Cover, 1999),

$$\begin{aligned} &H(X_t, X_{t+1}, \ldots, X_n \mid x_{<t}) \\ &\quad = H(X_t \mid x_{<t}) \\ &\qquad + H(X_{t+1}, \ldots, X_n \mid X_t, x_{<t}). \end{aligned}$$

The left-hand side is exactly $F(x_{<t})$ by definition (5). The first term on the right-hand side is the token entropy H_t (1). The second term is a

conditional entropy over the random variable X_t:

$$\begin{aligned} &H(X_{t+1}, \ldots, X_n \mid X_t, x_{<t}) \\ &= \sum_{v \in \mathcal{V}} p_t(v)\, H\big(X_{t+1}, \ldots, X_n \\ &\qquad \mid X_t{=}v,\, x_{<t}\big) \\ &= \sum_{v \in \mathcal{V}} p_t(v)\, F(x_{<t}, v) \\ &= \mathbb{E}_{V \sim p_t}\big[F(x_{<t}, V)\big] \\ &= \bar{F}_t. \end{aligned}$$

Combining the two sides gives

$$F(x_{<t}) = H_t + \bar{F}_t.$$

A.2 Proof of Proposition 1 (decomposition identity)

Starting from the definition of per-token future-entropy change,

$$\delta_t(x_t) = F(x_{<t}) - F(x_{\leq t}),$$

and substituting the chain-rule identity in Proposition 1:

$$\begin{aligned} \delta_t(x_t) &= \big(H_t + \bar{F}_t\big) - F(x_{\leq t}) \\ &= H_t + \big(\bar{F}_t - F(x_{\leq t})\big) \\ &= H_t + B_t(x_t). \end{aligned}$$

A.3 Proof of the zero-mean property of B_t and of the identity $\mathbb{E}[\delta_t(X_t)] = H_t$

Taking expectation under $X_t \sim p_t$:

$$\begin{aligned} \mathbb{E}_{X_t \sim p_t}\big[B_t(X_t)\big] &= \mathbb{E}_{X_t \sim p_t}\big[\bar{F}_t - F(x_{<t}, X_t)\big] \\ &= \bar{F}_t - \mathbb{E}_{X_t \sim p_t}\big[F(x_{<t}, X_t)\big] \\ &= 0. \end{aligned}$$

The second line uses linearity of expectation and the fact that $\bar{F}_t$ is a constant with respect to X_t. The last equality follows directly from the definition of $\bar{F}_t$ in (6).

Applying expectation to the decomposition in Proposition 1:

$$\begin{aligned} \mathbb{E}_{X_t \sim p_t}\big[\delta_t(X_t)\big] &= H_t + \mathbb{E}_{X_t \sim p_t}\big[B_t(X_t)\big] \\ &= H_t. \end{aligned}$$

A.4 Per-token future-entropy change can be negative

A common intuition is that "each additional token can only reduce uncertainty," which would suggest $\delta_t(x_t) \geq 0$ always. This intuition is incorrect for specific token realizations. The well-known inequality "conditioning reduces entropy" states

$$\mathbb{E}_{X_t \sim p_t}\big[F(x_{<t}, X_t)\big] \leq H(X_{t+1}, \ldots, X_n \mid x_{<t}).$$

which holds only *in expectation* over X_t, not pointwise. For a low-probability token x_t that opens a wide set of diverse continuations, $F(x_{\leq t})$ can exceed $F(x_{<t})$, yielding $\delta_t(x_t) < 0$.

Minimal example. Consider a binary vocabulary $\mathcal{V} = \{A, B\}$ at position 2 with $p_2(A) = 0.99$ and $p_2(B) = 0.01$. Suppose choosing A makes X_3 deterministic, $F(x_1, A) = 0$, while choosing B makes X_3 uniform, $F(x_1, B) = \ln 2 \approx 0.693$. Then $F(x_{<2}) = H_2 + \bar{F}_2 \approx 0.056 + 0.007 = 0.063$, and for the realized token B, $\delta_2(B) = 0.063 - 0.693 < 0$. The rare token B therefore *increases* future uncertainty rather than reducing it. From the credit-assignment perspective, $\delta_t(x_t) < 0$ identifies tokens that *diverge* the trajectory—a signal that is invisible when only the non-negative quantity H_t is used.

A.5 Proof of Proposition 2 (unbiased suffix-entropy estimator)

Proof. Fix a prefix $x_{\leq t}$ and consider the random suffix $(X_{t+1}, X_{t+2}, \ldots, X_n)$ sampled autoregressively from the model distribution $p(\cdot \mid x_{\leq t})$. By Equation (10),

$$\widehat{E}_t = \frac{1}{n-t} \sum_{s=t+1}^{n} h_s(X_{<s}),$$

where

$$h_s(x_{<s}) \triangleq -\sum_{v \in \mathcal{V}} p(v \mid x_{<s}) \log p(v \mid x_{<s}).$$

Note that h_s depends only on the prefix up to position $s-1$, and is itself a deterministic function of $(X_{t+1}, \ldots, X_{s-1})$ once $x_{\leq t}$ is fixed.

By linearity of expectation,

$$\mathbb{E}[\widehat{E}_t \mid x_{\le t}] = \frac{1}{n-t} \sum_{s=t+1}^{n} \mathbb{E}[h_s(X_{<s}) \mid x_{\le t}]. \tag{12}$$

For each s, the random variable $h_s(X_{<s})$ depends only on $(X_{t+1}, \ldots, X_{s-1})$. Writing $\xi = x_{t+1:s-1}$, the expectation is

$$\mathbb{E}[h_s(X_{<s}) \mid x_{\le t}] = \sum_{\xi} p(\xi \mid x_{\le t})\, h_s(x_{\le t}, \xi).$$

By definition of conditional entropy,

$$\begin{aligned} H(X_s \mid x_{\le t}) &= \sum_{\xi} p(\xi \mid x_{\le t})\, H(X_s \mid x_{\le t}, \xi) \\ &= \sum_{\xi} p(\xi \mid x_{\le t})\, h_s(x_{\le t}, \xi), \end{aligned}$$

where we used $H(X_s \mid x_{<s}) = h_s(x_{<s})$. Here and below, $H(X_s \mid x_{\le t})$ is shorthand for $H(X_s \mid X_{t+1:s-1}, x_{\le t})$: the conditional entropy of X_s averaged over the random intermediate tokens $X_{t+1:s-1}$, given the fixed prefix $x_{\le t}$.

Therefore $\mathbb{E}[h_s(X_{<s}) \mid x_{\le t}] = H(X_s \mid X_{t+1:s-1}, x_{\le t})$ for every $s \in \{t+1, \ldots, n\}$. Summing across s and applying the entropy chain rule (Cover, 1999),

$$\begin{aligned} &\sum_{s=t+1}^{n} H(X_s \mid X_{t+1:s-1},\, x_{\le t}) \\ &\quad = H(X_{t+1}, \ldots, X_n \mid x_{\le t}) \\ &\quad = F(x_{\le t}), \end{aligned}$$

where the final equality is the definition (5) of future entropy. Substituting back into (12) yields

$$\mathbb{E}[\widehat{E}_t \mid x_{\le t}] \;=\; \frac{F(x_{\le t})}{n-t}.$$

□

Off-distribution bias. If the suffix is drawn from a distribution $q(\cdot \mid x_{\le t}) \neq p(\cdot \mid x_{\le t})$—for example, from a stale policy π_{old} while the entropies h_s are computed under the current policy π_θ, or from a fixed dataset in a teacher-forcing evaluation—the same derivation yields

$$\begin{aligned} &\mathbb{E}_{X_{t+1:n} \sim q}[\widehat{E}_t \mid x_{\le t}] \\ &\quad = \frac{1}{n-t} \sum_{s=t+1}^{n} \sum_{x_{<s}} q(x_{<s} \mid x_{\le t})\, h_s^{(p)}(x_{<s}), \end{aligned}$$

where $h_s^{(p)}$ denotes the local entropy under the model p. The bias relative to $F(x_{\le t})/(n-t)$ is

$$\begin{aligned} \text{Bias}_t = \frac{1}{n-t} \sum_{s=t+1}^{n} \sum_{x_{<s}} \big(&q(x_{<s} \mid x_{\le t}) \\ &- p(x_{<s} \mid x_{\le t})\big)\, h_s^{(p)}(x_{<s}), \end{aligned} \tag{13}$$

which depends on the correlation between $h_s^{(p)}$ and the likelihood-ratio gap $q - p$ across prefixes. The sign of the bias is not determined a priori. In our GRPO setting, rollouts are drawn from the current policy (or a very recently updated one within the standard importance-weighting range), so $q \approx p$ at the positions used for credit modulation; Appendix B makes this statement explicit by bounding the distribution-shift component under a suffix-level KL condition. Because our method further consumes $\widehat{E}_t$ only through *ranks* within the batch-level high-entropy set $\mathcal{T}$, a uniform additive shift in $\widehat{E}_t$ across positions in $\mathcal{T}$ does not change the ranking and hence does not affect the modulator $m_t^{(k)}$.

Assumption 3 (Global old-current policy proximity)**.** Let $\mathcal{U}$ be the set of all sampled nonterminal token positions in the batch, and let $\mathcal{T} \subseteq \mathcal{U}$ be the high-entropy routed subset. For each position $i = (k, t) \in \mathcal{U}$, let $q_i = \pi_{\text{old}}(x_{>t} \mid x_{\le t}^{(k)})$ and $p_i = \pi_\theta(x_{>t} \mid x_{\le t}^{(k)})$ denote the old and current suffix distributions from the same prefix. The average suffix-level KL over the sampled positions is small:

$$\bar{\kappa}_{\mathcal{U}} \triangleq \frac{1}{|\mathcal{U}|} \sum_{i \in \mathcal{U}} D_{\text{KL}}(q_i \| p_i) \ll 1.$$

This is a policy-level proximity condition and does not impose a separate KL condition on the high-entropy subset $\mathcal{T}$.

B Off-policy low-tail stability of FEPO

This appendix makes the off-policy argument explicit. Rollouts are sampled from π_{old}, but suffix

entropies are scored under π_θ. The theory compares the *expected* scores $\bar{Z}^{q\to p}$ and Z^p; FEPO training uses a *single* realized suffix from π_{old} to form (10), so single-sample noise is not part of the deterministic proof and is checked in Section 5.3. We first show that if the old and current suffix distributions have small average KL on the sampled position set $\mathcal{U}$, then the global average induced shift in the expected future-entropy rate is bounded. We then use Markov's inequality, restrict the resulting bad set to $\mathcal{T}$, and apply a deterministic quantile argument to show that this bounded shift can change low-tail membership only at inherited large-drift positions or near the oracle β-quantile boundary.

Formal statement (full version of Proposition 3). The main text states the conclusion only; this appendix gives notation, bounds, and proof. Let $\mathcal{U}$ be all sampled nonterminal token positions and let $\mathcal{T} \subseteq \mathcal{U}$ be the routed high-entropy set, with density $\rho_{\mathcal{T}} \triangleq |\mathcal{T}|/|\mathcal{U}|$. For each $i = (k,t) \in \mathcal{U}$, let $p_i = \pi_\theta(\cdot \mid x^{(k)}_{\le t})$ and $q_i = \pi_{\text{old}}(\cdot \mid x^{(k)}_{\le t})$ be suffix distributions, and let Z^p_i and $\bar{Z}^{q\to p}_i$ be the fixed-normalization scores obtained from the true π_θ future-entropy rate and from the q_i-expected suffix entropy scored under π_θ, respectively. FEPO routing uses a single q_i-sampled suffix to form the realized score (10); the bounds below apply to the expectation-level scores $\bar{Z}^{q\to p}$ and Z^p, while single-suffix noise is assessed in Section 5.3. Let

$$\bar{\kappa}_{\mathcal{U}} \triangleq \frac{1}{|\mathcal{U}|}\sum_{i\in\mathcal{U}} D_{\text{KL}}(q_i \| p_i),$$
$$C_Z \triangleq c_{\text{norm}} \log |\mathcal{V}|.$$

Then the global average expected-score drift obeys

$$\frac{1}{|\mathcal{U}|}\sum_{i\in\mathcal{U}} \big|\bar{Z}^{q\to p}_i - Z^p_i\big| \le C_Z\sqrt{\bar{\kappa}_{\mathcal{U}}/2}. \tag{14}$$

For any threshold $\eta > 0$, define the global large-drift set and its restriction to $\mathcal{T}$:

$$\mathcal{B}^{\mathcal{U}}_\eta \triangleq \{i \in \mathcal{U} : |\bar{Z}^{q\to p}_i - Z^p_i| > \eta\},$$
$$\mathcal{B}^{\mathcal{T}}_\eta \triangleq \mathcal{B}^{\mathcal{U}}_\eta \cap \mathcal{T},$$
$$\varepsilon^{\mathcal{T}}_\eta \triangleq \frac{|\mathcal{B}^{\mathcal{T}}_\eta|}{|\mathcal{T}|}.$$

Markov's inequality gives a global bad-position fraction, and restriction to $\mathcal{T}$ gives

$$\varepsilon^{\mathcal{T}}_\eta \le \frac{1}{\rho_{\mathcal{T}}}\frac{C_Z}{\eta}\sqrt{\bar{\kappa}_{\mathcal{U}}/2}. \tag{15}$$

Let $Q_\gamma(Z^p)$ be the empirical γ-quantile of the oracle current-policy scores on $\mathcal{T}$, with γ clipped to $[0,1]$. If $\varepsilon^{\mathcal{T}}_\eta$ is comparable to β or $1-\beta$, the clipped bound remains valid but becomes nearly vacuous; the useful regime is $\varepsilon^{\mathcal{T}}_\eta \ll \min\{\beta, 1-\beta\}$. Define the expanded oracle boundary band

$$\mathcal{C}_\eta \triangleq \big\{i \in \mathcal{T} : Q_{\beta-\varepsilon^{\mathcal{T}}_\eta}(Z^p) - 2\eta \le Z^p_i,\; Z^p_i \le Q_{\beta+\varepsilon^{\mathcal{T}}_\eta}(Z^p) + 2\eta\big\}.$$

Then, up to the standard one-sample ambiguity of empirical quantiles,

$$L_\beta(\bar{Z}^{q\to p}) \triangle L_\beta(Z^p) \subseteq \mathcal{B}^{\mathcal{T}}_\eta \cup \mathcal{C}_\eta. \tag{16}$$

Thus the expected off-policy low tail can differ from the oracle current-policy low tail only on the large-drift positions inherited by $\mathcal{T}$ and on an expanded empirical boundary band. Training uses one realized suffix per position rather than $\bar{Z}^{q\to p}$, so the bound above is for the distribution-shift (expectation) term only.

Setup. Index the sampled positions in $\mathcal{U}$ by $i = (k,t)$, and write the fixed prefix as $s_i = x^{(k)}_{\le t}$ and the remaining length as $L_i = n^{(k)} - t$. Let

$$p_i(\cdot) = \pi_\theta(\cdot \mid s_i), \qquad q_i(\cdot) = \pi_{\text{old}}(\cdot \mid s_i)$$

denote the current-policy and old-policy suffix distributions from the same prefix, both viewed as distributions over the same L_i-token horizon (or equivalently their L_i-token truncations). For a suffix $y_{>t}$, define the current-policy entropy-rate functional

$$g^p_i(y_{>t}) \triangleq \frac{1}{L_i}\sum_{u=t+1}^{n^{(k)}} H_{\pi_\theta}(Y_u \mid s_i, y_{t+1:u-1}).$$

Since each token entropy is at most $\log|\mathcal{V}|$, we have $0 \le g^p_i \le G$ with $G = \log|\mathcal{V}|$. The true current-policy future-entropy rate and its off-policy expectation are

$$e^p_i \triangleq \mathbb{E}_{Y\sim p_i}[g^p_i(Y)],$$
$$\bar{e}^{q\to p}_i \triangleq \mathbb{E}_{Y\sim q_i}[g^p_i(Y)].$$

Let Z_i^p and $\bar{Z}_i^{q\to p}$ be obtained from e_i^p and $\bar{e}_i^{q\to p}$ using the same length-bucket affine normalization as in Definition 2. We treat the bucket statistics as fixed during routing and write

$$c_{\text{norm}} \triangleq \max_b \frac{1}{\sigma_b + \varepsilon}.$$

For a score vector Z, let $Q_\beta(Z)$ be its empirical β-quantile and

$$L_\beta(Z) \triangleq \{i : Z_i \le Q_\beta(Z)\}$$

be the low-tail set used by FEPO.

Distribution-shift bound. For each position i, because $g_i^p/G \in [0,1]$, the variational characterization of total variation gives

$$\begin{aligned}\left|\bar{e}_i^{q\to p} - e_i^p\right| &= \left|\mathbb{E}_{q_i} g_i^p - \mathbb{E}_{p_i} g_i^p\right| \\ &\le G\,\mathrm{TV}(q_i, p_i).\end{aligned}$$

Pinsker's inequality then yields

$$\left|\bar{e}_i^{q\to p} - e_i^p\right| \le G\sqrt{\frac{1}{2} D_{\mathrm{KL}}(q_i \| p_i)}.$$

Applying the fixed length-bucket normalization gives

$$\left|\bar{Z}_i^{q\to p} - Z_i^p\right| \le C_Z \sqrt{\frac{1}{2} D_{\mathrm{KL}}(q_i \| p_i)}.$$

Averaging over $i \in \mathcal{U}$ and using Jensen's inequality for the concave square-root function,

$$\frac{1}{|\mathcal{U}|} \sum_{i\in\mathcal{U}} \left|\bar{Z}_i^{q\to p} - Z_i^p\right| \le C_Z \sqrt{\bar{\kappa}_{\mathcal{U}}/2},$$

which proves (14). Markov's inequality then proves (15). Thus a global average suffix-level KL bound between π_{old} and π_θ controls the fraction of positions with large expected score drift, and the factor $1/\rho_{\mathcal{T}}$ accounts for restricting that global bad set to the routed subset.

Low-tail stability. By definition of $\mathcal{B}_\eta^{\mathcal{T}}$, every non-bad position in $\mathcal{T}$ has score drift at most η. This gives the following robust empirical-quantile bounds:

$$\begin{aligned}Q_\beta(\bar{Z}^{q\to p}) &\ge Q_{\beta-\varepsilon_\eta^{\mathcal{T}}}(Z^p) - \eta, \\ Q_\beta(\bar{Z}^{q\to p}) &\le Q_{\beta+\varepsilon_\eta^{\mathcal{T}}}(Z^p) + \eta,\end{aligned}$$

up to the usual one-sample ambiguity caused by ties. For the upper bound, at least a β fraction of positions in $\mathcal{T}$ are non-bad and have $Z_i^p \le Q_{\beta+\varepsilon_\eta^{\mathcal{T}}}(Z^p)$, hence their shifted scores are at most $Q_{\beta+\varepsilon_\eta^{\mathcal{T}}}(Z^p) + \eta$. For the lower bound, at most a $(\beta - \varepsilon_\eta^{\mathcal{T}})$ fraction of positions in $\mathcal{T}$ have $Z_i^p < Q_{\beta-\varepsilon_\eta^{\mathcal{T}}}(Z^p)$, and adding all bad positions contributes at most another $\varepsilon_\eta^{\mathcal{T}}$ fraction. Therefore fewer than a β fraction of shifted scores can fall below $Q_{\beta-\varepsilon_\eta^{\mathcal{T}}}(Z^p) - \eta$, which gives the stated lower quantile bound. When the empirical score density around $Q_\beta(Z^p)$ is bounded, choosing η to balance the inherited Markov term and the boundary-band width gives a total membership-change rate that vanishes with $\bar{\kappa}_{\mathcal{U}}$ for fixed $\rho_{\mathcal{T}} > 0$; for example, the natural scaling is $\eta \asymp (\sqrt{\bar{\kappa}_{\mathcal{U}}}/\rho_{\mathcal{T}})^{1/2}$ up to constants.

Now consider a non-bad position whose membership changes between the true current-policy low tail and the off-policy expected low tail. If $i \in L_\beta(Z^p)$ but $i \notin L_\beta(\bar{Z}^{q\to p})$, then

$$\begin{aligned}Z_i^p &\le Q_\beta(Z^p), \\ \bar{Z}_i^{q\to p} &> Q_\beta(\bar{Z}^{q\to p}).\end{aligned}$$

Using $\bar{Z}_i^{q\to p} \le Z_i^p + \eta$ and $Q_\beta(\bar{Z}^{q\to p}) \ge Q_{\beta-\varepsilon_\eta^{\mathcal{T}}}(Z^p) - \eta$ gives

$$Z_i^p > Q_{\beta-\varepsilon_\eta^{\mathcal{T}}}(Z^p) - 2\eta.$$

Together with $Z_i^p \le Q_\beta(Z^p)$, this puts i in the lower part of the expanded boundary band. The reverse case, $i \notin L_\beta(Z^p)$ but $i \in L_\beta(\bar{Z}^{q\to p})$, similarly implies $Z_i^p \le Q_{\beta+\varepsilon_\eta^{\mathcal{T}}}(Z^p) + 2\eta$. Adding the bad positions $\mathcal{B}_\eta^{\mathcal{T}}$ proves (16). Thus global old/current suffix-distribution shift can only affect the off-policy *expected* low-tail selection through a Markov-controlled bad set inherited by $\mathcal{T}$ and the empirical mass in an expanded band around the true current-policy boundary. The remaining risk beyond this deterministic bound is single-suffix sampling noise (realized Z versus $\bar{Z}^{q\to p}$).

Relation to policy updates. The KL quantity above is a global suffix-level average over sampled positions:

$$\bar{\kappa}_{\mathcal{U}} = \frac{1}{|\mathcal{U}|} \sum_{i \in \mathcal{U}} D_{\mathrm{KL}}\big(q_i(x_{>t}) \| p_i(x_{>t})\big).$$

By the chain rule for KL, define the stepwise divergence

$$d_u(x_{<u}) \triangleq D_{\mathrm{KL}}\big(\pi_{\mathrm{old}}(\cdot \mid x_{<u}) \| \pi_\theta(\cdot \mid x_{<u})\big).$$

Then the suffix-level KL decomposes as

$$D_{\mathrm{KL}}(q_i \| p_i) = \mathbb{E}_{q_i} \sum_{u=t+1}^{n^{(k)}} d_u(x_{<u}). \tag{17}$$

PPO-style small updates and KL regularization therefore provide a mechanism for keeping the global average $\bar{\kappa}_{\mathcal{U}}$ small on sampled prefixes, without requiring the high-entropy routed positions to satisfy a separate KL assumption. The empirical diagnostic below checks the resulting rank fidelity directly.

Empirical verification. Because FEPO ultimately depends on ranks and uses a single realized suffix rather than the off-policy expectation $\bar{e}^{q \to p}$, we assess the gap left by the proof in two ways. First, Section 5.3 reports low-q/large-B_t agreement against an MC=128 same-prefix reference, which probes whether smaller q more often matches the larger-gap direction. Second, Figure 6 reports the Spearman correlation between old-policy and current-policy suffix-entropy scores inside $\mathcal{T}$, computed per prompt and averaged across training. Spearman is a complementary *global* ranking-stability statistic: high correlation indicates that few pairwise orderings change; together with (16), this supports the rank-stability consequence of Assumption 3 in practice.

C Length-bucket statistics for FEPO

This appendix details how FEPO maintains the length-bucket statistics (μ_b, σ_b^2) used in the length-corrected suffix-entropy score (10).

Welford's online update. For each bucket b we maintain a running count N_b, a running mean μ_b, and a running second-moment accumulator $M_{2,b}$. For each in-bucket sample x (i.e. a value $\widehat{E}_t^{(k)}$ whose length bucket equals b), the update is

$$\begin{aligned} N_b &\leftarrow N_b + 1, \\ \delta_1 &\leftarrow x - \mu_b, \\ \mu_b &\leftarrow \mu_b + \delta_1 / N_b, \\ \delta_2 &\leftarrow x - \mu_b, \\ M_{2,b} &\leftarrow M_{2,b} + \delta_1\, \delta_2, \\ \sigma_b^2 &\leftarrow M_{2,b} / \max(N_b - 1,\, 1). \end{aligned} \tag{18}$$

Welford's algorithm is the standard numerically stable estimator of the running variance and avoids the catastrophic cancellation that arises in the naive mean-of-squares formulation.

Lagged use within a training step. At each training step, the routing decision in batch i uses only the statistics accumulated through batch $i-1$. The new samples in batch i are folded into the bucket statistics via Equation (18) only after the routing decision has been made. This lagged scheme guarantees that the modulator is independent of the samples it scores, eliminating an obvious source of self-referential bias.

EMA smoothing. To reduce step-to-step fluctuations, we additionally apply an exponential moving-average smoothing on top of the Welford estimates,

$$\begin{aligned} \mu_b^{\mathrm{EMA}} &\leftarrow (1-\eta)\, \mu_b^{\mathrm{EMA}} + \eta\, \mu_b, \\ (\sigma_b^2)^{\mathrm{EMA}} &\leftarrow (1-\eta)\, (\sigma_b^2)^{\mathrm{EMA}} + \eta\, \sigma_b^2, \end{aligned} \tag{19}$$

with a small EMA rate $\eta \in (0, 1]$. The score in (10) is then computed using $(\mu_b^{\mathrm{EMA}}, (\sigma_b^2)^{\mathrm{EMA}})$.

Warm-up and bucket fallback. When a bucket has too few samples, the variance estimator is unreliable. We therefore use the following deterministic fallback rule. Let $N_{\min}$ be a minimum sample-count threshold; for each query bucket b, FEPO returns the statistics of the smallest superset that has at least $N_{\min}$ samples, in the following order: (i) bucket b itself; (ii) the union of b and its two adjacent buckets $b-1, b+1$; (iii) coarser length buckets obtained by widening the bucket boundaries; (iv) the global pool over all lengths. During an initial

warm-up phase ($N_b < N_{\min}$ for all buckets), FEPO does not modulate any positions, and the modulator in (11) reduces to $m_t^{(k)} = 1$ everywhere; this disables FEPO's intervention until enough statistics have been collected.

Score clipping. Finally, the length-corrected score is clipped to a bounded range

$$Z_t^{(k)} \leftarrow \mathrm{clip}\big(Z_t^{(k)},\, -c,\, c\big), \tag{20}$$

with c a small constant (e.g. $c = 3$). Clipping limits the influence of bucket outliers; since FEPO consumes $Z_t^{(k)}$ only through the batch-level rank $q_t^{(k)}$, clipping has no effect on positions whose ranks are unambiguously inside or outside the low-tail set.

Pseudocode. The complete bucketed scoring procedure is summarized in Algorithm 1.

D Entropy-rate normalization

The future entropy $F(s)$ as defined in (5) is a *sum* over all remaining positions. When continuations have different lengths, the raw sum is dominated by length rather than per-token uncertainty, making cross-candidate comparisons unreliable.

We address this by normalizing to the *entropy rate*: for a continuation of length T following prefix s,

$$F^{\mathrm{rate}}(s) \;\triangleq\; \frac{F(s)}{T}.$$

Under the model's own distribution, the expected entropy rate equals the expected log-perplexity of the continuation (Cover, 1999):

$$\mathbb{E}\left[\frac{1}{T}\sum_{\tau=1}^{T}\big(-\log p(y_\tau \mid y_{<\tau})\big)\right] = \frac{1}{T}\sum_{\tau=1}^{T} H(Y_\tau \mid y_{<\tau}),$$

which is exactly $F^{\mathrm{rate}}(s)$ when the continuation is sampled from the model itself.

After normalization, F^{rate} and the token entropy H_t share the same unit (nats/token), making the bias term $\bar{F}_t^{\mathrm{rate}} - F^{\mathrm{rate}}(x_{\le t})$ directly comparable across positions and across candidates with different continuation lengths.

Algorithm 1 Length-bucket scoring for FEPO at training step i.

Input: batch of suffix-entropy rates $\{\widehat{E}_t^{(k)}\}$ with lengths $\{L_t^{(k)}\}$; cached statistics $(\mu_b^{\mathrm{EMA}}, (\sigma_b^2)^{\mathrm{EMA}}, N_b)$ from step $i-1$.
Output: length-corrected scores $\{Z_t^{(k)}\}$, updated statistics for step i.
for each (k, t) in the batch **do**
  $b \leftarrow b(L_t^{(k)})$
  $(\mu, \sigma^2, N) \leftarrow$ statistics of the smallest superset of b with $N \ge N_{\min}$
  if $N < N_{\min}$ (warm-up) **then**
    $Z_t^{(k)} \leftarrow 0$ ▷ disables FEPO modulation
  else
    $Z_t^{(k)} \leftarrow (\widehat{E}_t^{(k)} - \mu)/(\sqrt{\sigma^2} + \varepsilon)$
    $Z_t^{(k)} \leftarrow \mathrm{clip}(Z_t^{(k)}, -c, c)$
  end if
end for
Routing decision: compute batch-level ranks $\{q_t^{(k)}\}$ inside $\mathcal{T}$ and modulators $\{m_t^{(k)}\}$ from $\{Z_t^{(k)}\}$.
for each (k, t) in the batch **do**
  Update bucket $b(L_t^{(k)})$ statistics $(N_b, \mu_b, M_{2,b}, \sigma_b^2)$ via the Welford recursion (18).
end for
Apply EMA smoothing in (19) to update $(\mu_b^{\mathrm{EMA}}, (\sigma_b^2)^{\mathrm{EMA}})$.

E Implementation Details

Training uses 16×H800 with a global batch size of 128, a mini batch size of 64, a group size of 16, and a learning rate of 1×10^{-6}. For the optimizer we use the standard GRPO surrogate of (4) with a symmetric clip range $\epsilon_{\mathrm{clip}} = 0.2$, a low-variance KL penalty against π_{ref} with coefficient $\beta_{\mathrm{KL}} = 10^{-3}$, and a cosine learning-rate schedule with 10% warm-up. For generation, we use temperature 1.0, top-p 1.0, a maximum prompt length of 2048 tokens, and a maximum response length of 8192 tokens. The runtime summary is reported in Table 5.

F Artifacts and Data Documentation

We use existing scientific artifacts and cite their creators in Section 5.1: dapo17k for train-

Table 4: Artifacts and data used in this work. "Official eval split" means that we use the benchmark as released by its creators, without constructing a new split.

Artifact	Role	Statistics used here	Domain
dapo17k	train	∼17k prompts; 300-example probe subset	math reasoning
MATH 500	eval/probe	500 problems	math reasoning
AIME24/AIME25	eval	official eval split; 32 samples/prompt	contest math
AMC	eval	official eval split; 32 samples/prompt	contest math
OlympiadBench	eval	official eval split; 32 samples/prompt	olympiad math
Minerva	eval	official eval split; 32 samples/prompt	math reasoning
Qwen3-4B/8B	model	4B and 8B backbones	language model

ing (Yu et al., 2025); AIME24/AIME25 and AMC from the Mathematical Association of America (Mathematical Association of America, 2026a,b); MATH 500 (Lightman et al., 2023); OlympiadBench (He et al., 2024); Minerva (Lewkowycz et al., 2022); and the Qwen3-4B/Qwen3-8B backbones (Yang et al., 2025). The artifacts are used for academic research on mathematical reasoning and reinforcement-learning training signals. We use each third-party model and dataset under its original release terms, and we do not redistribute third-party datasets or modified third-party model weights as part of this submission. The accompanying code release is intended to support research reproducibility of FEPO-style training and evaluation, rather than deployment.

The language data used in this work consists of mathematical-reasoning prompts and solutions, primarily in English. We do not collect new human-subject data. Because the datasets are public mathematical problem collections rather than user-generated conversations or social-media text, we do not apply additional personally identifying information or offensive-content filtering beyond the checks and curation performed by the original dataset providers.

Table 4 summarizes the artifact coverage and statistics used in our experiments. For public benchmarks whose exact split sizes are defined by their original releases, we use the official evaluation splits without creating new train/dev/test partitions. All main evaluation numbers are reported with Mean@32 and Pass@32, i.e., 32 generated completions per prompt.

G Training dynamics and cost

This appendix summarizes training-time behavior complementing the diagnostics in Section 5.3. For each model (Qwen3-4B and Qwen3-8B) and each method (GRPO, GTPO, FEPO), we log three quantities at every step: (i) policy entropy H_t averaged over response tokens; (ii) policy KL loss; and (iii) mean response length. Across runs, FEPO neither accelerates nor prevents the standard slow decay of policy entropy, its KL-loss trajectory remains comparable to the baselines, and its response-length trajectory tracks GRPO closely. For FEPO we additionally log suffix-length, low-tail and high-head positive/negative ratios, low-entropy means inside $\mathcal{T}$, and suffix-rate percentiles; Figure 7 shows that these traces remain stable across training and are consistent with the sign-asymmetric modulation design.

Table 5: Training cost summary for GRPO and FEPO on Qwen3-4B and Qwen3-8B, reported as seconds per step (s/step) and total training time (h/total).

Model	Metric	GRPO	FEPO
Qwen3-4B	s/step	∼122	∼146
	h/total	∼23	∼26
Qwen3-8B	s/step	∼272	∼302
	h/total	∼51	∼57

Table 6: Ablations and hyperparameters. Average test accuracy is reported as a delta against the FEPO default on AIME24/AIME25.

Ablation	Setting	Avg.
α_{pos}	0.2	1.0
	0.3	-0.06
	0.4	-0.07
Entropy threshold	1.0	1.0
	0.8	-0.13
	top10%	-0.12
Length corr.	w/ bucket correction	1.0
	w/o bucket correction	-0.15
α_{neg}	0.2	1.0
	0.1	-0.02
	0.0	-0.12
	-0.2	-0.18
High-head	none	1.0
	$\alpha_{\text{high}} = -0.2$	-0.03
	$\alpha_{\text{high}} = +0.2$	-0.07
Tail fraction β	0.1	-0.13
	0.2	1.0
	0.3	-0.01
Cap ratio	no_cap	-0.1
	0.02cap	-0.01
	0.05cap	1.0
	0.1cap	-0.05

G.1 Ablation summary

Table 6 summarizes the major design sweeps on AIME24/AIME25.

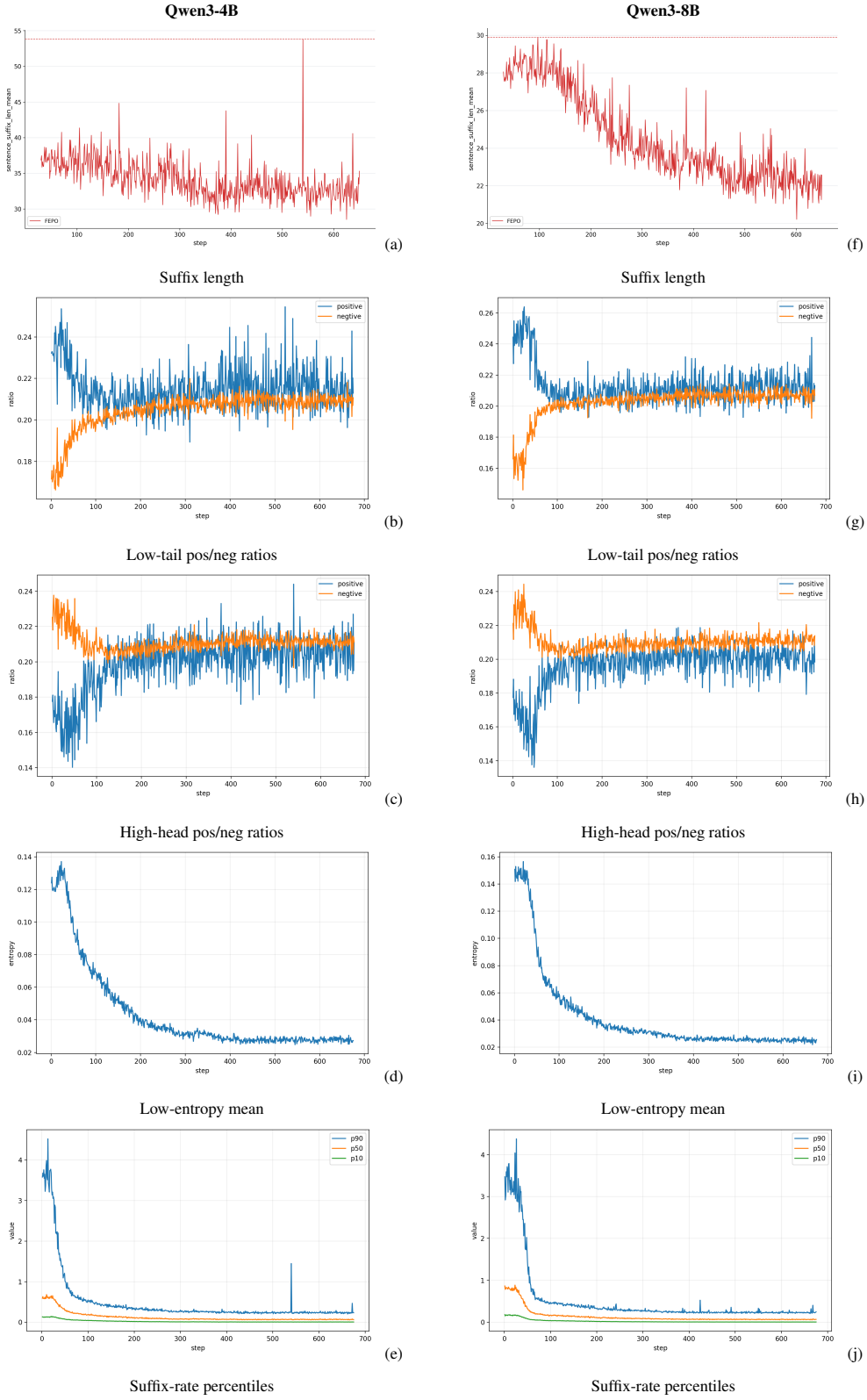

Figure 7: FEPO-specific training diagnostics. Left column: Qwen3-4B (a)–(e); right column: Qwen3-8B (f)–(j). From top to bottom: suffix length, low-tail positive/negative ratios, high-head positive/negative ratios, low-entropy mean, and suffix-rate percentiles.