REINFORCE PRO MAX: SIGN-DEPENDENT SCALING AND CAUSAL PREFIX MASKING

Anonymous authors
Paper under double-blind review

ABSTRACT

REINFORCE Pro Max combines sign-dependent advantage scaling and causal prefix masking for critic-free language-model reinforcement learning. We analyze how these transformations affect update directions, retained importance weights, and the relation between the training objective and expected reward. For Max, we derive the binary-reward coefficients and conditional local ascent, retaining response-length and sample-dependent scaling effects. For Pro, we bound the retained-weight second moment by an envelope times the accepted importance mass, plus a lower-side reentry term. This term accounts for accepted tokens following a lower-prefix excursion. An exact decomposition connects the scaled, masked importance-weighted objective to the terminal-reward surrogate. Applying existing Adaptive and Mixed bounds gives a sufficient reward-improvement condition with explicit scaling and masking costs. A two-step family has bounded retained second moments and a positive reward lower bound despite arbitrarily large unmasked second moments. A shared-parameter counterexample exhibits simultaneous objective ascent and reward decrease.

1 INTRODUCTION

Critic-free objectives such as RLOO and GRPO learn from terminal rewards without a value model (Ahmadian et al., 2024; Shao et al., 2024). In language-model training, however, a response-level reward produces many token updates. Response-level centering weights each response equally, whereas expanded token advantages weight responses by their active lengths. A second issue arises when rollout and actor policies differ: a per-token correction uses only the current ratio, while a whole-response mask makes the same decision for every position. These issues motivate REINFORCE Pro Max, which combines sign-dependent token normalization with causal prefix masking.

We analyze reward-improvement conditions with sampled RLOO groups, numerical safeguards, current-to-rollout ratios, and the active-token denominator.

Method and mathematical results. REINFORCE Max scales positive and negative advantages separately. Its ideal factors balance signed token mass and set the second moment to one. We derive its binary-reward coefficients and conditions under which the safeguarded update remains a positive multiple of the success-probability gradient. Variable within-class lengths and shared parameters across prompts introduce explicit covariance terms. The ideal binary rule equals matched token-weighted standardization; the safeguarded coefficients additionally depend on lengths and numerical regularization.

REINFORCE Pro selects each position using the current-to-rollout prefix geometric mean and retains a token-level importance coefficient. Its second-moment bound separates an envelope times the accepted importance mass from a lower-side reentry term. This term accounts for accepted tokens whose preceding prefix lies below the lower threshold.

Relation to prior theory. PPO constrains token ratios through clipping (Schulman et al., 2017). We instead use a masked importance-weighted objective and DPPO's terminal-reward decomposition (Qi et al., 2026) and TRM's Adaptive and Mixed bounds (Li et al., 2026) to connect this surrogate to reward. CTPO (Zhang et al., 2026b) uses cumulative ratios; MinPRO (Lei et al., 2026) clips and detaches the product of the current ratio and smallest preceding ratio. PNPO (Zhang

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.

et al., 2026a) uses current-to-behavior prefix geometric means as both weights and mask statistics. CPPO (Mao et al., 2026) masks token contributions using position-weighted divergence thresholds and cumulative prefix budgets. REINFORCE Pro uses current-to-rollout geometric means only for masking and retains a single token-level importance ratio. These choices motivate our retained-coefficient and reward-error analyses. The terminal RLOO signal depends on continuations and sibling responses; we integrate its baseline over complete prompt groups. Prefix change-of-measure identities apply directly to prefix-measurable functions; our outcome signals require the group conditioning argument. The normalization analysis of Ge et al. (2026) uses exact on-policy gradients scaled by population reward standard deviations. Here coefficients and token counts depend jointly on sampled responses. Length-only weighting tradeoffs (Ding et al., 2026) give a related perspective.

For the combined method, an exact objective decomposition isolates the scaling and masking costs. A fixed reference scale removes the distortion of matching common scaling under full retention. Applying the DPPO/TRM reward bounds gives a population improvement condition. In a two-step family, this bound stays positive and retained second moments remain bounded as unmasked moments grow without bound (Example 5.2). A shared-parameter counterexample shows that an increasing masked objective can accompany decreasing reward.

2 Finite model and the population reference

Prompts follow a fixed finite law. A policy is a normalized kernel on a finite token vocabulary and complete autoregressive histories $c_t = (x, y_{<t})$. Responses terminate by a fixed cap T. After EOS, all policies share an absorbing transition and the action mask is zero; the EOS-emitting action itself remains active. Rewards satisfy $|R(x, y)| \le R_{\max}$, and $J(q) = \mathbb{E}_{x,y\sim q} R(x, y)$. The rollout policy $p = \pi_{\text{roll}}$ and evaluated policy $q = \pi_\theta$ have mutual support at every context. At a represented token, define the single ratio

$$\rho_t(q) = \frac{q(y_t \mid c_t)}{p(y_t \mid c_t)}.$$

Both the prefix statistic and the token coefficient use this ratio. We analyze the masked surrogate without PPO clipping or a separate importance-correction multiplier. Rewards are sparse and terminal; KL, entropy, and auxiliary-loss coefficients are zero. The rollout law and sampled advantages remain fixed during an update, whereas the mask is recomputed from the current policy and detached during differentiation.

Raw surrogate and reward error. Let b be independent of its own response conditional on the prompt, with the rollout and baseline laws fixed during evaluation. The centered raw surrogate is

$$\mathcal{S}_p(q) = \mathbb{E}_p\left[(R - b)\sum_t m_t(\rho_t - 1)\right] = \mathbb{E}_p\left[R\sum_t m_t(\rho_t - 1)\right].$$

Conditional ratio normalization cancels the baseline, and the tower property identifies this with the per-step advantage surrogate (Proposition A.4). Normalization, masking, and finite-batch reduction enter separately through the objective decomposition below.

Write $\bar{D}_t = \mathbb{E}_{x,c_t\sim p}\,\mathrm{TV}(p(\cdot \mid c_t), q(\cdot \mid c_t))$. Assume local TV and $D_{\mathrm{KL}}(p\|q)$ are bounded by ε and δ at every context before T, respectively. Let $\bar{D}_{\text{seq}} = \mathbb{E}_x\,\mathrm{TV}(P_x^p, P_x^q)$ for the complete response laws. Combining the Adaptive and Mixed bounds of Li et al. (2026), as proved in Theorem A.5, gives

$$\begin{aligned}
|J(q) - J(p) - \mathcal{S}_p(q)| &\le B(p, q),\\
B(p, q) &= \min\{B_{\text{Adap}}, B_{\text{Mix}}\},\\
B_{\text{Adap}} &= 4R_{\max}\sum_{t=1}^{T}\bar{D}_t\min\{1, (T-t)\varepsilon, \sqrt{(T-t)\delta/2}\},\\
B_{\text{Mix}} &= 4R_{\max}T\min\{1, \varepsilon\}\bar{D}_{\text{seq}}.
\end{aligned}$$

Adaptive controls conditional continuation TV; Mixed controls marginal prefix shifts. Both use policy-level bounds at all contexts before T, distinct from sampled mask decisions. Section A proves both bounds and derives their DPPO and Pinsker–Marginal specializations.

3 REINFORCE MAX

Max scales positive and negative advantages by separate positive factors α and β. In this paper, its inputs are RLOO signals $A_i = R_i - (n-1)^{-1}\sum_{j\neq i} R_j$ for $n \geq 2$ responses to one prompt. Sparse terminal reward with $\gamma = 1$ assigns A_i to every active token. Let $P = \{(i,t) : A_i > 0, m_{i,t} = 1\}$ and $N = \{(i,t) : A_i < 0, m_{i,t} = 1\}$. For nonempty P, N, we choose $\alpha, \beta > 0$ as the ideal factors balancing signed token mass and setting the second moment to one:

$$\sum_P \alpha A_i + \sum_N \beta A_i = 0, \qquad \frac{1}{|P| + |N|}\left(\sum_P (\alpha A_i)^2 + \sum_N (\beta A_i)^2\right) = 1.$$

The normalized signal is $H_{i,t} = \alpha A_i$ on P, βA_i on N, and zero on raw zero tokens. Consequently signs and within-sign ordering are preserved while positive and negative token mass are balanced. Numerical safeguards introduce the moment residuals in Proposition B.6, retained in the objective decomposition.

Closed-form factors. For nonempty P and N, let $S_+ = \sum_P A_i > 0$, $S_- = -\sum_N A_i > 0$, and $Q_\pm$ be the corresponding sums of squared advantages. Writing $m = |P| + |N|$, the two constraints give

$$\kappa = \frac{S_+}{S_-}, \qquad \alpha = \sqrt{\frac{m}{Q_+ + \kappa^2 Q_-}}, \qquad \beta = \kappa\alpha.$$

Zero signed mass gives $\alpha S_+ = \beta S_-$; the second-moment constraint then determines the unique positive solution. Positive factors preserve individual signs but can change the summed gradient direction. For an exact RLOO group, $\sum_i A_i = 0$, whereas the expanded token sum is $\sum_i T_i A_i$. Unequal response lengths weight each advantage differently; the asymmetric factors balance this token mass before importance weighting and multiplication by score vectors.

Proposition 3.1 (Binary RLOO normalization and response-length balance)**.** *Suppose a prompt group contains $k \geq 1$ responses with scalar reward r_+ and $m \geq 1$ with reward $r_- < r_+$, and write $n = k + m$ and $\Delta = r_+ - r_- > 0$. Work in the RLOO branch and use the ideal factors in Eq.* (11)*. Let P and N be the total active-token counts in the higher- and lower-reward responses, respectively; assume $P, N > 0$. Then the raw RLOO advantages are a_+ and $-a_-$, where*

$$a_+ = \frac{m\Delta}{n-1}, \qquad a_- = \frac{k\Delta}{n-1}.$$

Every active token in the two classes has normalized advantage

$$\hat{A}_+ = \alpha a_+ = \sqrt{N/P}, \qquad \hat{A}_- = -\beta a_- = -\sqrt{P/N}. \tag{1}$$

Writing $\bar{T}_+ = P/k$ and $\bar{T}_- = N/m$ for the class-conditional mean response lengths, the factor ratio and aggregate coefficient balance satisfy

$$\frac{\beta}{\alpha} = \kappa = \frac{Pa_+}{Na_-} = \frac{\bar{T}_+}{\bar{T}_-}, \qquad P\hat{A}_+ = -N\hat{A}_-. \tag{2}$$

These conclusions concern scalar advantages before importance weighting, prefix masking, or multiplication by score vectors.

Proof of Proposition 3.1. Subtracting the sibling mean from each reward gives $a_+ = m\Delta/(n-1) > 0$ and $a_- = k\Delta/(n-1) > 0$. The length-weighted first and second sums are Pa_+, Na_- and Pa_+^2, Na_-^2. Hence

$$\kappa = \frac{Pa_+}{Na_-}, \qquad D = Pa_+^2 + \kappa^2 Na_-^2 = a_+^2 \frac{P(P+N)}{N}.$$

Substituting into $\alpha = \sqrt{(P+N)/D}$ and $\beta = \alpha\kappa$ gives $\alpha a_+ = \sqrt{N/P}$ and $\beta a_- = \sqrt{P/N}$. Also $\kappa = Pm/(Nk) = (P/k)/(N/m) = \bar{T}_+/\bar{T}_-$, and $P\sqrt{N/P} = N\sqrt{P/N}$ proves aggregate balance. □

Update direction under class-dependent lengths. Fix one prompt with rewards $R \in \{0,1\}$ and $n \geq 2$ iid on-policy responses. Let $s(\theta) = \Pr_{p_\theta}(R=1)$, $g = \nabla s(\theta)$ and $u_\theta(y) = \nabla \log p_\theta(y)$. Assume class-dependent lengths $0 < L_0, L_1 \leq L_{\max}$, accepted prefix masks. With the safeguarded factors of Proposition B.8, lower clamp $\eta \in (0,1]$, and group token count L, the response-score coefficient is $C_i = H_i/L$. The expected update satisfies

$$U := \mathbb{E}\left[\sum_i C_i(Y_{1:n}) u_\theta(Y_i)\right] = \Lambda g, \qquad \Lambda \geq \eta / L_{\max} > 0.$$

To see this, condition on the $n-1$ siblings Z of response i. The coefficient then depends on Y_i only through its reward label. The zero-mean score identity gives the multiplier $\Lambda = \sum_i \mathbb{E}_Z[C_i(1,Z) - C_i(0,Z)]$. The sibling mean lies in $[0,1]$, each positive scale is at least η, and every group has at most $nL_{\max}$ active tokens. Each conditional contrast is therefore at least $\eta/(nL_{\max})$; summing proves the bound. The optional uniform replacement adds a nonnegative contrast on all-success groups and zero on all-failure groups, preserving this bound.

Ideal normalization removes any common positive scaling of the raw advantages. Hence mean-baseline and RLOO signals coincide after ideal normalization; binary signals also equal matched token standardization (Proposition B.4 and Corollary B.5).

When the two scales change the update direction. The SGA analysis of Li & Liu (2025b, Section 2) separates gradient alignment from stochastic noise. For Max, the relevant first-order quantity is the inner product of the reward gradient with the expected scaled update. The following decomposition identifies how dependence between the scale and the response affects that alignment.

Proposition 3.2 (Within-class length covariance decomposition)**.** *Drop the class-dependent-length premise and let $C_i(y_{1:n})$ be the actual response-score coefficient produced by the implementation, including its length-weighted normalization, total-token denominator and optional uniform-scale term. For a fixed direction v, write $f_v(y) = \langle v, u_\theta(y)\rangle$. Assume both binary response classes have positive finite probability. Define Λ_C as the sum over coordinates and sibling draws of the difference between the conditional class means of C_i, and define $\mathcal{E}_v$ as the corresponding weighted within-class covariance between C_i and f_v. Then the exact identity is*

$$\mathbb{E}\left[\sum_i C_i(Y_{1:n}) f_v(Y_i)\right] = \Lambda_C\, \mathbb{E}[\mathbf{1}_{\{R=1\}} f_v(Y)] + \mathcal{E}_v,$$
$$|\mathcal{E}_v| \leq \sqrt{V_C\, n\, V_{f,v}}.$$

Here, writing Z for the $n-1$ independent sibling responses and inserting Y at coordinate i, define $\bar{C}_i(q,Z) := \mathbb{E}[C_i(Y,Z) \mid R(Y)=q, Z]$ and $\bar{f}_v(q) := \mathbb{E}[f_v(Y) \mid R(Y)=q]$. The variances are

$$V_C := \sum_i \mathbb{E}_{Z,Y}[(C_i(Y,Z) - \bar{C}_i(R(Y),Z))^2],$$
$$V_{f,v} := \mathbb{E}_Y[(f_v(Y) - \bar{f}_v(R(Y)))^2].$$

For $v = g = \nabla s(\theta)$, this gives $\langle g, U\rangle \geq \Lambda_C \|g\|^2 - \sqrt{V_C\, n\, V_{f,g}}$. A positive right-hand side implies $s(\theta + tU) > s(\theta)$ for all sufficiently small $t > 0$ (Corollary B.9).

Proof. Condition on the sibling responses and split the current response support into the two reward classes. Subtracting and adding the conditional class means of both C_i and f_v gives the displayed identity; the terms with one centered factor vanish within each class. Applying the finite Cauchy–Schwarz inequality gives the square-root bound. The score identity $\mathbb{E}[\mathbf{1}_{\{R=1\}} f_g(Y)] = \|g\|^2$ yields the ascent bound; differentiability gives local improvement. Class-dependent lengths make $\mathcal{E}_v = 0$, recovering Proposition B.8. □

Multivalued rewards distinguish the two rules. For rewards $(0, 2/3, 1)$ and active lengths $(1,1,4)$, the actual RLOO advantages are $(-5/6, 1/6, 2/3)$, with token mean $1/3$ and variance $11/36$. Token standardization changes the second response's coefficient from positive to negative, since $1/6 - 1/3 = -1/6$; sign-dependent scaling keeps $\alpha/6 > 0$. The sign change distinguishes token standardization from sign-dependent scaling for multivalued rewards.

From a prompt direction to population ascent. For a fixed prompt with binary success probability $s_x(\theta)$, write $g_x = \nabla s_x(\theta)$ and $U_x = \Lambda_x g_x + e_x$ for the conditional population update. With the stated on-policy, accepted-mask conditions, the covariance analysis bounds the projection of e_x separately from the positive gradient multiplier Λ_x. Across a fixed finite prompt law, let $G = \mathbb{E}_x g_x$, $U = \mathbb{E}_x U_x$, and $f_x = \langle G, g_x \rangle$. Then Corollary B.10 gives

$$\langle G, U \rangle \geq \mathbb{E}_x \Lambda_x \|G\|^2 - \sqrt{\mathrm{Var}_x(\Lambda_x)\,\mathrm{Var}_x(f_x)} - \mathbb{E}_x |\langle G, e_x \rangle|.$$

A strictly positive right-hand side implies local reward improvement along U in the differentiable finite model. Differing positive prompt scales can otherwise reweight conflicting gradients. The two subtracted terms quantify the effects of between-prompt reweighting and within-class coefficient variation, respectively.

4 REINFORCE Pro: Causal Prefix Mask

4.1 Masked surrogate objective

Following the masked objectives of DPPO and TRM (Qi et al., 2026; Li et al., 2026), Pro uses

$$S_{\mathrm{Pro}}(q) = \sum_{i,t} \omega_{i,t} M_{i,t}(q) \rho_{i,t}(q) A_i, \qquad \rho_{i,t}(q) = \frac{q(y_{i,t} \mid c_{i,t})}{p(y_{i,t} \mid c_{i,t})}.$$

Here $\omega_{i,t} = m_{i,t} / \sum_{j,s} m_{j,s}$ includes all active tokens, including rejected ones. Pro Max replaces A_i by the scaled signal $H_{i,t}$.

Prefix masking criterion. TRM assigns one mask to an entire response; Pro makes a separate decision at each position using only its observed prefix. Define

$$g_{i,t}(q) = \exp\left(\frac{\sum_{s \leq t} m_{i,s} \log \rho_{i,s}(q)}{\max(1, \sum_{s \leq t} m_{i,s})} \right), \qquad M_{i,t}(q) = \mathbf{1}\{\ell_- \leq g_{i,t}(q) \leq \ell_+\}.$$

For $0 < \ell_- < 1 < \ell_+$, the geometric mean selects contributions; $\rho_{i,t}$ supplies their token weight. Later prefixes may restore acceptance without changing earlier decisions. DPPO instead uses a divergence- and advantage-dependent mask. Signals and masks are detached during differentiation; ρ carries the gradient (Appendix D).

Exact objective decomposition. Let k enumerate represented active positions, with nonnegative reduction weights ω_k, raw signal A_k, scaled signal H_k, and $M_k = M_k(q)$, $\rho_k = \rho_k(q)$. Define

$$U = \sum_k \omega_k \rho_k A_k, \qquad G = \sum_k \omega_k \rho_k (1 - M_k) A_k,$$
$$N = \sum_k \omega_k \rho_k M_k (H_k - A_k).$$

Proposition 4.1 (Finite-batch decomposition)**.** *The masked objective satisfies*

$$-\mathcal{L} = U - G + N, \qquad |-\mathcal{L} - U| \leq |G| + |N|.$$

The identity holds for every realized batch, including random response lengths and policy-dependent masks.

Here G and N record rejection and scaling. Adding and subtracting the retained raw contribution proves the identity (Proposition D.1). This adapts the discarded-contribution analysis (Li & Liu, 2025a, Section 1.3) to a tokenwise prefix mask.

4.2 A bound on retained-weight moments

The quantity to control is the retained coefficient $M_t \rho_t$. Individual token ratios can be arbitrarily large under prefix-average acceptance. For a normalized two-action, two-step policy pair, the ratios on one path can be (K^{-1}, K), so $M_2 = 1$ for any thresholds containing one, yet

$$\mathbb{E}_{\pi_{\mathrm{roll}}}[(M_2 \rho_2)^2] \geq \frac{K^3}{(K+1)^2} \geq K/4 \quad (K \geq 1).$$

Algorithm 1 REINFORCE Pro Max: core transformations.

Require: Batch $(y_i, A_i, m_{i,t})$ with prompt-group index $b(i)$; policies p, q; Pro interval $[\ell_-, \ell_+]$, step size η

1: $(\alpha_b, \beta_b) \leftarrow \mathrm{MaxScales}(A_b, m_b)$ for each prompt group b
2: $H_{i,t} \leftarrow \mathrm{stopgrad}\big(A_i(\alpha_{b(i)}\mathbf{1}_{A_i>0} + \beta_{b(i)}\mathbf{1}_{A_i<0})\big)$ {Max}
3: $z_{i,t} \leftarrow \log q(y_{i,t} \mid c_{i,t}) - \log p(y_{i,t} \mid c_{i,t})$; $\rho \leftarrow \exp z$
4: $g_{i,t} \leftarrow \exp\left(\dfrac{\sum_{s\le t} m_{i,s} z_{i,s}}{\max(1, \sum_{s\le t} m_{i,s})}\right)$
5: $M \leftarrow \mathrm{stopgrad}(\mathbf{1}_{\ell_- \le g \le \ell_+})$ {Pro}
6: $\mathcal{L} \leftarrow -\dfrac{\sum_{i,t} m_{i,t} M_{i,t} \rho_{i,t} H_{i,t}}{\sum_{i,t} m_{i,t}}$
7: $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}$

Proposition C.3 gives the construction. Controlling this moment requires accounting for reentry into the acceptance interval. The mask bounds a prefix average rather than each token ratio. The reward bound assumes local TV and KL conditions on the policy pair.

Second-moment bound with lower-side reentry. For this bound, the rollout and current kernels have mutual support, and M_t denotes the mask on the full padded tree prefix. Let $W_t = \prod_{s\le t} \rho_s$, $W_0 = 1$, $\tau_+ = \log \ell_+$ and $\tau_- = -\log \ell_-$. The lower-excursion contribution is $\mathcal{R}_t^- = \mathbb{E}_{\pi_{\mathrm{roll}}}[\mathbf{1}\{\log W_{t-1} < -(t-1)\tau_-\} M_t \rho_t^2]$. If $\log W_{t-1} \ge -(t-1)\tau_-$ and the current token is accepted, then $\rho_t \le C_t = e^{t\tau_+ + (t-1)\tau_-}$. Hence

$$\mathbb{E}_{\pi_{\mathrm{roll}}}[(M_t\rho_t)^2] \le C_t \mathbb{E}_{\pi_{\mathrm{roll}}}[M_t\rho_t] + \mathcal{R}_t^- \le C_t + \mathcal{R}_t^-.$$

Proof, step 1: a local envelope. Acceptance gives $\log W_t \le t\tau_+$; subtracting the previous log-product bounds ρ_t. *Step 2: conditional normalization.* At a history h above the lower threshold, we have

$$\mathbb{E}_{\pi_{\mathrm{roll}}}[M_t\rho_t^2 \mid h] \le C_t \mathbb{E}_{\pi_{\mathrm{roll}}}[M_t\rho_t \mid h] \le C_t \sum_y q(y \mid h) = C_t.$$

Step 3: split the histories. Controlled histories contribute at most $C_t \mathbb{E}_{\pi_{\mathrm{roll}}}[M_t\rho_t]$; the remaining histories contribute $\mathcal{R}_t^-$. Normalization bounds the first moment by one. Here $\mathbb{E}_{\pi_{\mathrm{roll}}}[M_t\rho_t]$ is importance-weighted retained mass. The argument adapts the Seq-MIS envelope analysis (Li & Liu, 2025a, Section 1.3) to the current-policy prefix mask.

For $T\tau_+ \le a$, $T\tau_- \le b$, $C_t \le e^{a+b}$ for $t \le T$. Upper-side reentry gives $\rho_t \le e^{\tau_+}$; lower-side reentry contributes the additional moment term. The same bound controls the active-token coefficient $m_t M_t \rho_t$ (Theorem C.4).

Example C.5 gives a nontrivial regime: the mask retains half the rollout mass at every depth with $\mathcal{R}_t^- = 0$ and second moment $1/2$, while the unmasked first-token moment is arbitrarily large.

Causal rejection. The prefix log-average updates by a convex combination, allowing later ratios to restore acceptance (Lemma C.6). Early mismatch can suppress later terms; a deviation at t^* leaves earlier decisions intact. Theorem C.7 quantifies the rejection duration.

Algorithm 1 uses per-prompt Max scales and an active-token mean. Token-count weighting preserves this objective; the implementation's response-count weighting is treated in Appendix D.2.

$\mathrm{MaxScales}$ uses the two factors derived above, with the numerical safeguards in Proposition B.6; a fallback returns $(1, 1)$. The optional homogeneous-group replacement in Appendix B.5 bypasses this core transformation.

5 From the masked objective to reward improvement

Following DPPO and TRM, we bound the raw reward error, quantify the scaling and masking cost, and combine both in a reward-improvement bound.

5.1 Reward decomposition and sequence error

For a fixed candidate q and rollout policy p, first isolate the difference between true reward and the raw surrogate. We use DPPO's product decomposition (Qi et al., 2026, Appendix B.1–B.2) and TRM's factor-by-factor Adaptive proof (Li et al., 2026, Appendix D). *Step 1: exact reward decomposition.* At a fixed prompt, telescoping gives

$$J_x(q) - J_x(p) - \mathcal{S}_{p,x}(q) = \mathbb{E}_p\left[R \sum_t (\rho_t - 1)\left(\prod_{j>t} \rho_j - 1\right)\right].$$

J_x and $\mathcal{S}_{p,x}$ are prompt-conditional reward and raw surrogate. *Step 2: bound the continuation factor.* Conditioning on $c_{t+1} = (x, y_{\le t})$ fixes the token factor and gives continuation TV:

$$\mathbb{E}_p\left[\left|\prod_{j>t} \rho_j - 1\right| \mid c_{t+1}\right] = 2\,\mathrm{TV}(P^p_{>t}, P^q_{>t}).$$

Both continuation laws condition on that same prefix. The sequence-TV inequality bounds this quantity through subsequent local divergences:

$$\mathrm{TV}(P^p_{>t}, P^q_{>t}) \le \sum_{j>t} \mathbb{E}_{p(\cdot|c_{t+1})}\,\mathrm{TV}(p(\cdot \mid c_j), q(\cdot \mid c_j)) \le (T-t)\varepsilon.$$

To obtain the first inequality, telescope the two continuation probability products and sum absolute differences. Each candidate suffix sums to one, leaving rollout-weighted local TVs. Combining this with the unit TV bound and KL chain rule followed by Pinsker gives $b_{T-t} = \min\{1, (T-t)\varepsilon, \sqrt{(T-t)\delta/2}\}$. *Step 3: aggregate the error.* The tower property and $\mathbb{E}_{x,y\sim p}|\rho_t - 1| = 2\bar{D}_t$ give B_{Adap}.

Mixed bound: control the prefix distribution instead. Following Li et al. (2026, Appendix C.2), let d^p_t, d^q_t be the prefix laws at a fixed prompt and $g_t(c) = \mathbb{E}_{a\sim q(\cdot|c)} A^p_t(c,a)$, where $A^p_t(c,a) = \mathbb{E}_p[R \mid c, a] - \mathbb{E}_p[R \mid c]$. The performance-difference identity gives

$$J_x(q) - J_x(p) - \mathcal{S}_{p,x}(q) = \sum_t \big(\mathbb{E}_{d^q_t} g_t - \mathbb{E}_{d^p_t} g_t\big).$$

Conditional normalization gives $\|g_t\|_\infty \le 2R_{\max}\min(1,\varepsilon)$. Marginalization gives $\mathrm{TV}(d^p_t, d^q_t) \le \mathrm{TV}(P^p_x, P^q_x)$. Applying $|\mathbb{E}_P g - \mathbb{E}_Q g| \le 2\|g\|_\infty \mathrm{TV}(P,Q)$, then summing and averaging over prompts, gives B_{Mix}. Both routes bound the same error, so their minimum yields $B(p,q)$ (Appendix A.3).

5.2 Scaling and masking error

The rollout law, sampled RLOO signals and Max factors are fixed when comparing the current policy q with p. The mask is evaluated at each policy: $M_t(p) = 1$, since all reference ratios equal one. For one group, write $L = \sum_{i,t} m_{i,t} > 0$, $S_{\mathrm{PM}} = -\mathcal{L}$, and $\Delta S_{\mathrm{PM}} = S_{\mathrm{PM}}(q) - S_{\mathrm{PM}}(p)$. Group objectives enter the batch objective with weights proportional to their active-token counts. For a fixed reference scale $\lambda > 0$, define

$$D_\lambda(q) = \frac{1}{L}\sum_{i,t} m_{i,t}\left[|\rho_{i,t} - 1|\,|\lambda A_i - H_{i,t}| + \rho_{i,t}(1 - M_{i,t})|H_{i,t}|\right].$$

The first term measures scaling discrepancy along the policy increment; the second charges contributions rejected by the current-policy mask. The scale λ fixes comparison units and does not change the algorithm. In particular, common scaling $H = \lambda A$ removes the first term, and full retention removes the second. Both vanish at $q = p$. Detaching M during backpropagation does not make it constant between candidate policies; this comparison includes its change explicitly.

5.3 Population reward bound

Theorem 5.1 (Population reward improvement)**.** *Assume the finite autoregressive model and the support and divergence conditions of Theorem A.5. For each prompt draw $n \ge 2$ conditionally iid*

rollout responses and form their RLOO signals. Use the same sampled advantages, scaling factors and active-token reduction at p and q, and compute each mask from its own policy ratio. With expectation over the prompt and its response group, for any fixed $\lambda > 0$ let

$$\mathfrak{M}_\lambda(q) = \frac{1}{\lambda n}\mathbb{E}\big[L\big(\Delta S_{\mathrm{PM}} - D_\lambda(q)\big)\big].$$

Then

$$J(q) - J(p) \geq \mathfrak{M}_\lambda(q) - B(p,q). \tag{3}$$

Thus $\mathfrak{M}_\lambda(q) > B(p,q)$ suffices for strict reward improvement.

Proof. Step 1: recover the raw increment. Let $Z = \sum_i A_i \sum_t m_{i,t}(\rho_{i,t} - 1)$. Since $M(p) = 1$, the pointwise residual is

$$\lambda Z - L\Delta S_{\mathrm{PM}} = \sum_{i,t} m_{i,t}\big[(\rho_{i,t} - 1)(\lambda A_i - H_{i,t}) + \rho_{i,t}(1 - M_{i,t})H_{i,t}\big].$$

Its absolute value is at most $LD_\lambda(q)$, so $L(\Delta S_{\mathrm{PM}} - D_\lambda(q))/\lambda \leq Z$ on every group.

Step 2: integrate the group signal. Conditional ratio normalization gives $\mathbb{E}_p[\sum_t m_t(\rho_t - 1) \mid x] = 0$. Each sibling baseline is independent of its own response given x, so

$$\frac{1}{n}\mathbb{E}[Z \mid x] = \mathbb{E}_{y\sim p(\cdot|x)}\left[R(x,y)\sum_t m_t(\rho_t - 1)\right].$$

Averaging over prompts identifies $\mathbb{E}[Z]/n = \mathcal{S}_p(q)$. *Step 3: subtract the sequence error.* The combined bound yields $J(q) - J(p) \geq \mathcal{S}_p(q) - B(p,q) \geq \mathfrak{M}_\lambda(q) - B(p,q)$. □

For $K \geq 1$ iid prompt groups, let $L_\Sigma = \sum_b L_b$ and, for $F \in \{\Delta S_{\mathrm{PM}}, D_\lambda\}$, define $F_{\mathrm{batch}} = \sum_b L_b F_b / L_\Sigma$. Then the same margin is

$$\mathfrak{M}_\lambda(q) = \frac{1}{\lambda K n}\mathbb{E}\big[L_\Sigma\big(\Delta S_{\mathrm{PM,batch}} - D_{\lambda,\mathrm{batch}}\big)\big].$$

Multiplying by L_Σ restores the group contributions; iid averaging cancels K. The random denominator stays inside the expectation. If $M = 1$ and $H = \lambda A$, then $D_\lambda = 0$. Write $\mathfrak{M} = \mathfrak{M}_1$. The examples specify their reference scale and use the full rollout law, including rejected branches.

5.4 A POSITIVE REGIME AND A FAILURE BOUNDARY

Example 5.2 (Positive margin with nontrivial prefix rejection)**.** Fix $a = 1 - \zeta$, $0 < \zeta < 1/5$, $\tau = \log(5/4)$ and $B > e^{2\tau}$. On three tokens, let

$$p_1 = \left(a, \frac{\zeta}{B+1}, \frac{\zeta B}{B+1}\right), \qquad q_{\delta,1} = \left(a, \frac{\zeta B}{B+1}, \frac{\zeta}{B+1}\right).$$

The rollout continuation is uniform. The current policy is uniform after tokens $1, 2$, and after token 0 it is $q_\delta(\cdot \mid 0) = (1/3 + \delta, 1/3 - \delta, 1/3)$, where $0 \leq \delta \leq 1/15$. Use mask interval $[4/5, 5/4]$. The first ratios are $(1, B, B^{-1})$; on branch 0, second ratios are $(1 + 3\delta, 1 - 3\delta, 1)$. The mask therefore accepts exactly branch 0 at both positions. Set $R(y) = \mathbf{1}\{y_1 = y_2 = 0\}$. Then $J(q_\delta) - J(p) = a\delta$.

Use two iid rollout responses per group, four-token reduction $L = 4$, no group masking or `uniform_scale`, stabilizer $0 \leq \epsilon \leq 1/2$, product cap 10^8, and factor clamp $[\epsilon, 10]$. Mixed-reward groups have $H = s_\epsilon A$, where $s_\epsilon = \sqrt{4/(4+\epsilon)} \in [3/4, 1]$; equal-reward groups have zero signals. All rejected tokens remain in L. For $f_q(y) = \sum_t M_t(q)\rho_t(q)$, independence of the sibling response gives

$$\mathbb{E}[S_{\mathrm{PM}}(q)] = \frac{s_\epsilon}{2}\operatorname{Cov}_p(R, f_q), \qquad S_{\mathrm{PM}}(p) = 0.$$

Here $\mathbb{E}_p R = a/3$, $\mathbb{E}_p f_q = 2a$, and $\mathbb{E}_p[R f_q] = a(2 + 3\delta)/3$. Consequently,

$$\mathbb{E}[\Delta S_{\mathrm{PM}}] = s_\epsilon(a\zeta/3 + a\delta/2), \qquad \mathbb{E}[D_{s_\epsilon}] = s_\epsilon a\zeta/3, \qquad \mathfrak{M}_{s_\epsilon} = a\delta.$$

For the distortion, the scaling term vanishes. A rejected response has reward zero, expected absolute RLOO signal $a/3$, and total rejected ratio mass 2ζ, which gives the displayed expectation.

The root TV is $\bar{D}_1 = \zeta(B-1)/(B+1) < \zeta$; continuation TVs are at most δ. For $\zeta \le \delta$, all local TVs are at most δ, so the Adaptive bound gives $B_{\mathrm{Adap}} \le 4\bar{D}_1\delta$. Thus $\zeta = 1/100$ and $\delta = 1/30$ yield, uniformly in $B > 25/16$,

$$J(q_\delta) - J(p) \ge \mathfrak{M}_{s_\epsilon} - B(p,q) \ge (1-5\zeta)\delta = \frac{19}{600} > 0.$$

At depths one and two the retained second moments are a and $a(1+6\delta^2)$, with no lower-side reentry. They remain bounded as $\mathbb{E}_p[\rho_1^2] = a+\zeta(B^2+B^{-1})/(B+1)$ grows without bound. For a parameter update within this mask region, $\partial_\delta \mathbb{E} S_{\mathrm{PM}} = s_\epsilon a/2$ and $\partial_\delta J = a > 0$: the population ascent direction increases reward.

Proposition 5.3 (Objective ascent can lower reward under a shared parameter)**.** *There is a normalized, strictly positive two-step current-policy family with one shared parameter and a fixed rollout law such that arbitrarily small positive parameter changes increase the expected ProMax objective while decreasing reward. The mask has nontrivial prefix rejection.*

Proof. Let $p_1 = (1/2, 1/6, 1/3)$ with uniform continuations, and let $q_{\delta,1} = (1/2, 1/3, 1/6)$. Set

$$q_\delta(\cdot \mid 0) = (1/3+\delta, 1/3-\delta, 1/3), \qquad q_\delta(\cdot \mid 1) = (1/3-3\delta, 1/3+3\delta, 1/3),$$

with uniform continuation after token 2 and $0 \le \delta \le 1/45$. Use $[4/5, 5/4]$ for the prefix mask and $R(y) = \mathbf{1}\{y_2 = 0\}$. At the first token, the ratios are $(1, 2, 1/2)$. On branch 0, the second prefix products lie in $[14/15, 16/15]$, inside the squared mask interval. On branch 1 they are at least $2(1-9\delta) \ge 8/5 > 25/16$; on branch 2 they equal $1/2 < 16/25$. Thus both depths accept exactly branch 0 throughout this parameter interval.

With the same two-response groups and safeguards as above, $H = s_\epsilon A$. The full rollout law remains fixed. The accepted ratio sum changes by $3\delta \mathbf{1}_{y_1=0}(\mathbf{1}_{y_2=0} - \mathbf{1}_{y_2=1})$. Its mean is zero and its covariance with R is $\delta/2$. Therefore

$$\mathbb{E}[S_{\mathrm{PM}}(q_\delta) - S_{\mathrm{PM}}(q_0)] = s_\epsilon\delta/4 > 0, \qquad J(q_\delta) - J(q_0) = -\delta/2 < 0. \tag{4}$$

Here q_0 is the zero-parameter member of the same current-policy family, not an additional importance-sampling reference. The negative reward change follows from $\frac{1}{2}\delta - \frac{1}{3}(3\delta)$. The rejected branch carries the larger adverse change. This example separates local objective ascent from the sufficient population margin, which also accounts for discarded contributions and sequence error. □

6 Limitations

The guarantees apply to finite autoregressive models with bounded terminal rewards and the stated support, divergence, and reduction conditions. Local ascent concerns on-policy population directions and requires positive class masses and the covariance margins. The population reward bound compares the current and rollout policies with sampled signals held fixed; it includes the change of the prefix mask and the distortion from scaling. Its applicability depends on the resulting margin relative to the sequence-error bound.

7 Conclusion

REINFORCE Pro Max combines sign-dependent RLOO scaling with causal prefix masking. Its analysis connects sample-dependent scales to update alignment and retained importance mass to a second-moment bound with lower-side reentry. The reward bound restores the token normalization and subtracts scaling and rejection costs from the population objective increment. The smaller Adaptive or Mixed remainder then bounds the residual reward error. A two-step family has bounded retained second moments, a positive reward bound, and arbitrarily large unmasked second moments. The shared-parameter counterexample shows that objective ascent alone can lower reward. Both the mask and the retained token weight compare the current policy directly with rollout.

AI USE STATEMENT

Generative AI tools assisted manuscript preparation, including drafting, language polishing, and improving clarity and organization. They also supported literature comparison, mathematical analysis, and proof development and implementation. The authors reviewed the resulting text and technical content, checked the proofs and their correspondence with the implementation, and take responsibility for the final manuscript.

ETHICS STATEMENT

This work uses no human subjects or private user data.

REPRODUCIBILITY STATEMENT

The main text and appendix provide the theorem statements, assumptions, and proofs. The supplement contains Lean 4.19.0 proof sources, pinned dependencies, a statement-to-proof index, and an axiom audit. Its implementation reference files specify the loss, normalization, and reduction conventions of the analyzed objective. Instructions reproduce the formal checks without GPU training.

REFERENCES

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting REINFORCE style optimization for learning from human feedback in LLMs. *arXiv preprint arXiv:2402.14740*, 2024.

Fei Ding, Yongkang Zhang, Yuhao Liao, Zijian Zeng, and Huiming Yang. On the impossibility of unbiased and length-invariant policy optimization with outcome rewards, 2026. URL https://arxiv.org/abs/2607.23364.

Cheng Ge, Caitlyn Heqi Yin, Hao Liang, and Jiawei Zhang. Why grpo needs normalization: A local-curvature perspective on adaptive gradients, 2026. URL https://arxiv.org/abs/2601.23135.

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

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

Yingru Li and Jiacai Liu. Part 3: Trust region optimization via sequence masking. Technical blog post, November 2025a. URL https://richardli.xyz/post/rl-collapse-part3/.

Yingru Li and Jiacai Liu. Part 1: Why off-policy breaks RL—an SGA analysis framework. Technical blog post, October 2025b. URL https://richardli.xyz/post/rl-collapse-part1/.

Yingru Li and Jiacai Liu. Part 2: Applying the SGA framework—token v.s. sequence-level correction. Technical blog post, October 2025c. URL https://richardli.xyz/post/rl-collapse-part2/.

Yingru Li, Jiacai Liu, Jiawei Xu, Yuxuan Tong, Ziniu Li, Qian Liu, and Baoxiang Wang. Trust region masking for long-horizon LLM reinforcement learning. arXiv preprint arXiv:2512.23075v4, 2026. URL https://arxiv.org/abs/2512.23075v4.

Renjie Mao, Xiangxin Zhou, Lvfang Tao, Yixin Ding, Yu Shi, Yongguang Lin, Yuheng Wu, Honglin Zhu, Qian Qiu, and Wenxi Zhu. Beyond uniform token-level trust region in LLM reinforcement learning, 2026. URL https://arxiv.org/abs/2606.10968v2.

Penghui Qi, Xiangxin Zhou, Zichen Liu, Tianyu Pang, Chao Du, Min Lin, and Wee Sun Lee. Rethinking the trust region in LLM reinforcement learning, 2026. URL https://arxiv.org/abs/2602.04879v1.

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

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

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

Wenhao Zhang, Yibo Xie, Rui Wang, Jiahua Yang, Lei Jiang, Zibo Yang, Yawei Wang, Jiali Xu, jasperawang, Haoyang Long, Huan Xiong, and alantzhao. Reusing rollouts under policy lag: Prefix-normalized policy optimization for LLM reinforcement learning, 2026a. URL https://arxiv.org/abs/2608.01418v1.

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

A Finite model and reward reference

The proofs use a common finite autoregressive model: prompts and tokens have finite support, terminal rewards are bounded, and a shared absorbing EOS transition pads responses to horizon T. The EOS-emitting action is active; subsequent padding has action mask zero. Write $p = \pi_{\mathrm{roll}}$ and $q = \pi_\theta$. On active rollout support,

$$\rho_t = q(y_t \mid c_t)/p(y_t \mid c_t). \tag{5}$$

$$\log \rho_t = \log q(y_t \mid c_t) - \log p(y_t \mid c_t). \tag{6}$$

The action mask is predictable from the history. All reductions use its active positions. The analyzed objective uses one current-to-rollout ratio, a detached prefix mask, no PPO clipping, no KL or entropy term, and sparse terminal reward. Let d_t^p denote the rollout law of the history before action t. For finite distributions, $D_{\mathrm{TV}}(p,q) = \frac{1}{2}\sum_a |p_a - q_a|$ and $D_{\mathrm{KL}}(p\|q) = \sum_{a:p_a>0} p_a \log(p_a/q_a)$.

Assumption A.1 (Mutual Policy Support)**.** For all contexts c_t and tokens y_t in the vocabulary,

$$\pi_{\mathrm{roll}}(y_t|c_t) > 0 \quad \Longleftrightarrow \quad \pi_\theta(y_t|c_t) > 0.$$

On this common support, ρ_t and $\log \rho_t$ are well-defined, and conditional normalization gives $\mathbb{E}_{\pi_{\mathrm{roll}}}[\rho_t \mid c_t] = 1$. Tokens with zero mass under both policies contribute zero to rollout expectations and KL/TV sums; their ratios are set to 1. The common support may be a proper subset of the vocabulary.

Definition A.2 (Uniform Token-Level Divergence Bounds and Expected Per-Position TV)**.** For the policy pair $(\pi_{\mathrm{roll}}, \pi_\theta)$, let $\epsilon, \delta \geq 0$ denote finite uniform upper bounds whenever the result invoking them assumes such bounds:

$$D_{\mathrm{TV}}^{\mathrm{tok}}(c_t; \pi_{\mathrm{roll}}, \pi_\theta) \leq \epsilon,$$
$$D_{\mathrm{KL}}^{\mathrm{tok}}(c_t; \pi_{\mathrm{roll}}, \pi_\theta) \leq \delta, \qquad \text{for all relevant } t, c_t.$$

The KL direction is $\pi_{\mathrm{roll}}\|\pi_\theta$. The relevant contexts include all histories before the common horizon in the model under consideration, uniformly over its prompts. On the finite context set, mutual support makes the divergences finite, and their maxima provide uniform bounds. Define the rollout-weighted expected per-position TV by

$$\bar{D}_t := \mathbb{E}_{c_t \sim d_t^{\pi_{\mathrm{roll}}}}[D_{\mathrm{TV}}^{\mathrm{tok}}(c_t; \pi_{\mathrm{roll}}, \pi_\theta)].$$

Samples are generated from a behavior policy π_{roll} that generally differs from π_θ. Following Kakade & Langford (2002), the standard surrogate objective uses the per-step advantage under π_{roll},

$$A_t^{\pi_{\text{roll}}}(c_t, y_t) := \mathbb{E}_{\pi_{\text{roll}}}\big[R(x,y) \mid c_t, y_t\big] - \mathbb{E}_{\pi_{\text{roll}}}\big[R(x,y) \mid c_t\big].$$

This yields the per-step surrogate

$$\mathcal{L}_{\pi_{\text{roll}}}(\pi_\theta) = \sum_{t=1}^{T} \mathbb{E}_{c_t \sim d_t^{\pi_{\text{roll}}}}\Big[m_t(c_t)\, \mathbb{E}_{y_t \sim \pi_{\text{roll}}(\cdot|c_t)}\big[\rho_t\, A_t^{\pi_{\text{roll}}}(c_t, y_t)\big]\Big], \tag{7}$$

where $m_t(c_t) \in \{0,1\}$ is the predictable action mask from Definition A.3. In the unpadded fixed-horizon model $m_t \equiv 1$; with EOS padding it removes post-termination contexts, whose shared absorbing transition has zero advantage and unit ratio. The ratio ρ_t is defined in Eq. (5).

Definition A.3 (Trajectory-Level Surrogate)**.** Let b be a baseline that is independent of the sampled trajectory y conditional on the prompt x (e.g., a prompt-level baseline computed from other independent rollout samples in the same prompt group). The rollout policy and the baseline sampling law are fixed during the update. Let $m_t = m_t(c_t) \in \{0,1\}$ denote the predictable action mask: it equals one while generation is active, including the action that emits EOS, and zero at subsequent padding positions. Define the trajectory-level advantage

$$A := R(x,y) - b.$$

The unclipped, unnormalized trajectory-level surrogate underlying critic-free implementations is

$$\tilde{\mathcal{L}}_{\pi_{\text{roll}}}(\pi_\theta) := \mathbb{E}_{\pi_{\text{roll}}}\Big[A \cdot \sum_{t=1}^{T} m_t \rho_t\Big].$$

Proposition A.4 (Exact Centered Surrogate Equivalence)**.** *For finite prompt, action, and continuation supports, under mutual policy support and the finite/EOS-padded horizon convention, assume that the expectations below exist and that b and m_t satisfy Definition A.3. Then, for any current policy π_θ,*

$$\tilde{\mathcal{L}}_{\pi_{\text{roll}}}(\pi_\theta) - \tilde{\mathcal{L}}_{\pi_{\text{roll}}}(\pi_{\text{roll}}) = \mathcal{L}_{\pi_{\text{roll}}}(\pi_\theta) = \mathbb{E}_{\pi_{\text{roll}}}\left[R(x,y) \sum_{t=1}^{T} m_t(\rho_t - 1)\right]. \tag{8}$$

The subtracted quantity is constant during the policy update; the identity holds for every candidate policy with the stated support.

Equality averages over response and baseline laws. Random-length batch reductions are treated in Section D.

A.1 Centered Surrogate Equivalence and Error Decomposition

Proof of Proposition A.4. Write $Q_t(c_t, a) = \mathbb{E}_{\pi_{\text{roll}}}[R \mid c_t, a]$, $V_t(c_t) = \mathbb{E}_{\pi_{\text{roll}}}[R \mid c_t]$, and $\bar{b}(x) = \mathbb{E}[b \mid x]$. Conditional independence of b and policy normalization give

$$\mathbb{E}_{\pi_{\text{roll}}}[(R-b)\rho_t \mid c_t] = \sum_a \pi_\theta(a \mid c_t)(Q_t(c_t,a) - V_t(c_t)) + V_t(c_t) - \bar{b}(x).$$

The last two terms are independent of π_θ. At $\pi_\theta = \pi_{\text{roll}}$ the advantage sum is zero. Multiplying by the predictable mask $m_t(c_t)$, averaging over rollout contexts, and summing over t proves the first equality. The second follows by replacing b by zero and using $\mathbb{E}_{\pi_{\text{roll}}}[R(\rho_t - 1) \mid c_t] = \sum_a \pi_\theta(a \mid c_t)(Q_t(c_t,a) - V_t(c_t))$. On post-EOS padding states the advantage is zero and the ratio is one. □

The argument integrates both the continuation law and the independent baseline law at each context. Random-length batch reductions are treated in Proposition D.1 and Section D.

A.2 Population reward bound

We use the Adaptive and Mixed bounds of Li et al. (2026) in the finite model above, with bounded signed rewards and an average over prompts. The proof records the terminal-reward and EOS correspondence needed for the subsequent Max/Pro objective comparison.

Theorem A.5 (Finite-tree reward error bound)**.** *For a finite prompt distribution and finite vocabulary, let both policies be normalized history-dependent kernels satisfying Assumption A.1 and the fixed/EOS-padded horizon convention. Suppose $|R| \le R_{\max}$ with $R_{\max} \ge 0$, and let ϵ, δ bound the local TV and reverse KL at every context before horizon T, as in Definition A.2. Then*

$$B^*_{\mathrm{Adap}} = 4R_{\max} \sum_{t=1}^{T} \bar{D}_t \cdot \min\Big(1,\ (T-t)\,\epsilon,\ \sqrt{\tfrac{(T-t)\,\delta}{2}}\Big), \tag{9}$$

and, writing P^p_x, P^q_x for the complete response laws,

$$B_{\mathrm{Mix}} = 4R_{\max} T \min(1,\epsilon)\mathbb{E}_x\, \mathrm{TV}(P^p_x, P^q_x), \qquad B = \min\{B^*_{\mathrm{Adap}}, B_{\mathrm{Mix}}\}.$$

Then

$$|J(\pi_\theta) - J(\pi_{\mathrm{roll}}) - \mathcal{L}_{\pi_{\mathrm{roll}}}(\pi_\theta)| \le B, \qquad J(\pi_\theta) - J(\pi_{\mathrm{roll}}) \ge \mathcal{L}_{\pi_{\mathrm{roll}}}(\pi_\theta) - B.$$

A.3 Reward decomposition and sequence error

Proof of Theorem A.5. We combine DPPO's product decomposition (Qi et al., 2026, Appendix B.1) with TRM's conditional-TV and marginal-prefix arguments (Li et al., 2026, Appendices D and C.2). Proposition A.4 incorporates RLOO centering and EOS. *Step 1: exact reward decomposition.* Fix a prompt and write $p = \pi_{\mathrm{roll}}$, $q = \pi_\theta$. On rollout support, the trajectory likelihood ratio is $\prod_{t=1}^{T} \rho_t$. Telescoping this product gives

$$\prod_{t=1}^{T} \rho_t - 1 = \sum_{t=1}^{T} (\rho_t - 1) \prod_{j>t} \rho_j.$$

Changing measure and subtracting the raw surrogate therefore yields

$$J(q) - J(p) - \mathcal{S}_p(q) = \mathbb{E}_p\left[R \sum_{t=1}^{T} (\rho_t - 1) \left(\prod_{j>t} \rho_j - 1 \right) \right].$$

Step 2: bound the continuation factor. Condition on the full prefix $c_{t+1} = (x, y_{\le t})$, which fixes $\rho_t - 1$. The remaining product is the likelihood ratio of the two conditional continuation laws $P^q_{>t}(\cdot \mid c_{t+1})$ and $P^p_{>t}(\cdot \mid c_{t+1})$. Consequently,

$$\mathbb{E}_p\left[\left| \prod_{j>t} \rho_j - 1 \right| \middle|\, c_{t+1} \right] = 2 D_{\mathrm{TV}}(P^p_{>t}, P^q_{>t}).$$

For $n = T - t$ remaining actions, normalization bounds this TV by one. The conditional sequence-TV argument of Qi et al. (2026, Appendix B.2) gives the sharper step

$$\mathrm{TV}(P^p_{>t}, P^q_{>t}) \le \sum_{j=t+1}^{T} \mathbb{E}_{p(\cdot|c_{t+1})}\, \mathrm{TV}\big(p(\cdot \mid c_j), q(\cdot \mid c_j)\big) \le n\epsilon.$$

To obtain the first inequality, telescope the two products of conditional probabilities, using rollout factors before position j and candidate factors after it. The triangle inequality separates each local difference; summing the candidate continuation gives one, leaving the rollout-weighted local TVs. The kernels may depend on the complete history. The KL chain rule and Pinsker's inequality, as used by Li et al. (2026, Appendix D), give $\sqrt{n\delta/2}$. Thus it is at most $b_n = \min(1, n\epsilon, \sqrt{n\delta/2})$ at every continuation history. *Step 3: aggregate the error.* The tower property and $|R| \le R_{\max}$ give

$$\begin{aligned} |J(q) - J(p) - \mathcal{S}_p(q)| &\le 2R_{\max} \sum_{t=1}^{T} b_{T-t} \mathbb{E}_p |\rho_t - 1| \\ &= 4R_{\max} \sum_{t=1}^{T} b_{T-t} \bar{D}_t. \end{aligned}$$

The last equality uses conditional ratio normalization at each context. All divergence bounds concern contexts before T; conditioning in the tower step retains the dependence of the continuation on the current action. Averaging over prompts proves the stated bound and its one-sided version. On shared absorbing EOS transitions $\rho_t = 1$, so padding contributes neither a surrogate term nor a local divergence.

Alternative bound: marginal prefix shift. Follow the argument of Li et al. (2026, Appendix C.2). Write d_t^p, d_t^q for the laws of the full prefix and $g_t(c_t) = \mathbb{E}_{a\sim q(\cdot|c_t)} A_t^p(c_t, a)$. The performance-difference identity yields

$$J(q) - J(p) - \mathcal{S}_p(q) = \sum_t \big(\mathbb{E}_{d_t^q} g_t - \mathbb{E}_{d_t^p} g_t\big).$$

Since $|Q_t^p| \le R_{\max}$, conditional normalization gives $|g_t| \le 2R_{\max}\min(1, \epsilon)$. Total variation contracts under marginalization, giving $\mathrm{TV}(d_t^p, d_t^q) \le \mathrm{TV}(P_x^p, P_x^q)$. Applying the bounded-function expectation inequality at each position, summing over t, and averaging over prompts gives B_{Mix}. Both bounds control the same absolute error, hence their minimum is valid. The Mixed bound uses marginal prefix laws; the Adaptive bound controls continuation laws conditional on each history and current action. □

Relation to DPPO and TRM bounds. Using $\bar{D}_t \le \epsilon$ with $b_{T-t} \le (T-t)\epsilon$, or using $b_{T-t} \le 1$, recovers DPPO's two bounds (Qi et al., 2026, Appendix B.2–B.3):

$$B_{\mathrm{Adap}} \le \min\left\{2R_{\max}T(T-1)\epsilon^2, 4R_{\max}\sum_t \bar{D}_t\right\}.$$

Pinsker gives $\bar{D}_t \le \sqrt{\delta/2}$; integrating $\sqrt{x}$ bounds $\sum_{k<T}\sqrt{k} \le 2T^{3/2}/3$, yielding $B_{\mathrm{Adap}} \le (4/3)R_{\max}\delta T^{3/2}$, TRM's Pinsker–Marginal rate (Li et al., 2026). Retaining $\bar{D}_t$ avoids charging the largest divergence at every position. For remaining length $h = T - t$ and $\epsilon > 0$, $h\epsilon \le \sqrt{h\delta/2}$ for $0 < h \le \delta/(2\epsilon^2)$; beyond this crossover the KL factor is smaller. Both capped factors equal one if both reach one, and vanish at $h = 0$. If $\epsilon = 0$, the TV factor is zero throughout.

TRM's complete-sequence KL variant follows from the Mixed proof (Li et al., 2026, Appendix C.2). Let $K_x = D_{\mathrm{KL}}(P_x^p \| P_x^q)$ and $\bar{K} = \mathbb{E}_x K_x$. Applying Pinsker to the local advantage shift and full-response TV gives

$$\begin{aligned}|J(q) - J(p) - \mathcal{S}_p(q)| &\le B_{\mathrm{Mix}}^{\mathrm{KL}},\\ B_{\mathrm{Mix}}^{\mathrm{KL}} &:= 4R_{\max}T\min(1, \sqrt{\delta/2})\mathbb{E}_x \min(1, \sqrt{K_x/2})\\ &\le 4R_{\max}T\min(1, \sqrt{\delta/2})\min(1, \sqrt{\bar{K}/2}).\end{aligned}$$

The last step uses concavity of $u \mapsto \min(1, \sqrt{u/2})$. With inactive caps the last expression is $2R_{\max}T\sqrt{\delta\bar{K}}$, useful when $\bar{K} \ll T\delta$. This KL alternative need not improve the exact-TV bound. Their minimum with B_{Adap} can replace B in Theorem 5.1; scaling and rejection costs remain in the margin, and policy-divergence bounds remain assumptions.

B REINFORCE MAX: CONSTRUCTION AND CONDITIONAL EFFECTIVENESS

Max's defining operation is separate positive scaling of positive and negative advantages. The balance and second-moment constraints determine the factors; the update analysis below uses RLOO inputs. The score-function and independent-baseline identities are standard policy-gradient tools (Williams, 1992; Ahmadian et al., 2024). Conditioning on sibling responses retains the dependence of each factor and the token denominator on the sampled group.

B.1 RLOO AND TERMINAL-REWARD EXPANSION

Definition B.1 (Leave-One-Out Baseline)**.** Assume $n \ge 2$. For sample i, the baseline and shaped reward are:

$$b_i = \frac{1}{n-1}\sum_{j\ne i} r_j, \qquad \tilde{r}_i = r_i - b_i. \tag{10}$$

With sparse terminal reward, zero KL coefficient and $\gamma = 1$, each active token has the undiscounted signal $A_{i,t} = A_i = R_i - b_i$. The response scores satisfy $u_\theta(y) = \sum_t m_t \nabla_\theta \log p_\theta(y_t \mid c_t)$ by differentiating the finite autoregressive product.

Remark B.2 (Raw RLOO baseline identities)**.** Fix a prompt x, parameters in $\mathbb{R}^d$, and a fixed finite response support. In a neighborhood of the evaluated parameter, assume that $p_\theta(y \mid x)$ is positive, differentiable and normalized, and that $R(x, y)$ is parameter-independent. For $n \geq 2$ iid responses, set $f_i = \nabla \log p_\theta(y_i \mid x)$, $r_i = R(x, y_i)$, $b_i = (n-1)^{-1} \sum_{j \neq i} r_j$, and $\hat{g} = n^{-1} \sum_i (r_i - b_i) f_i$. Then

$$\mathbb{E}[\hat{g} \mid x] = \nabla J_x(\theta), \qquad J_x(\theta) = \mathbb{E}_{p_\theta(\cdot|x)} R(x, y).$$

For $\bar{r} = n^{-1} \sum_i r_i$, the mean-baseline estimator satisfies $n^{-1} \sum_i (r_i - \bar{r}) f_i = (n-1)\hat{g}/n$ on every group. These statements concern the raw response-averaged estimator, before normalization, masking or a random token denominator.

Proof. Differentiating finite sums gives $\mathbb{E}[f_i \mid x] = 0$ and $\mathbb{E}[r_i f_i \mid x] = \nabla J_x$. The sibling baseline b_i is independent of (r_i, f_i) given x, so $\mathbb{E}[b_i f_i \mid x] = 0$; average over i. The scaling identity follows from $r_i - \bar{r} = (n-1)(r_i - b_i)/n$.

□

B.2 Token normalization

All token sets below contain active positions only.

Definition B.3 (Adaptive Normalization)**.** For one prompt group, let $\mathcal{P} = \{(i,t) : A_{i,t} > 0\}$ and $\mathcal{N} = \{(i,t) : A_{i,t} < 0\}$, restricted to active tokens. Set $\hat{A} = \alpha A$ on $\mathcal{P}$, $\hat{A} = \beta A$ on $\mathcal{N}$, and $\hat{A} = 0$ on zero-advantage tokens. For nonempty sign classes, write

$$S_+ = \sum_{\mathcal{P}} A > 0, \qquad S_- = -\sum_{\mathcal{N}} A > 0,$$
$$Q_+ = \sum_{\mathcal{P}} A^2, \qquad Q_- = \sum_{\mathcal{N}} A^2, \qquad d = |\mathcal{P}| + |\mathcal{N}|.$$

The ideal factors $\alpha, \beta > 0$ enforce zero signed token mass and unit second moment over these d nonzero tokens:

$$\alpha S_+ - \beta S_- = 0, \qquad \alpha^2 Q_+ + \beta^2 Q_- = d.$$

Their unique positive solution is

$$\kappa = \frac{S_+}{S_-}, \qquad \alpha = \sqrt{\frac{d}{Q_+ + \kappa^2 Q_-}}, \qquad \beta = \kappa\alpha. \tag{11}$$

The first constraint fixes $\beta/\alpha = \kappa$; substitution into the second fixes α^2, and positivity selects its positive square root. For exact RLOO with positive active lengths, an empty sign class implies all advantages are zero. Such groups use the identity fallback in Appendix B.4; the two moment constraints apply to groups with both signs present. The implemented stabilizer, caps and clamps are treated separately in Proposition B.6.

Proposition B.4 (Scale invariance of ideal normalization)**.** *Fix a finite active-token set with both advantage signs present and use the ideal normalization of Definition B.3. For every $c > 0$, replacing all advantages by cA leaves every normalized advantage unchanged. Consequently, for any nondegenerate group with $n \geq 2$, applying the same ideal asymmetric normalization to the mean-baseline and RLOO signals gives identical token coefficients, including with unequal response lengths:*

$$A_i^{\text{mean}} = \frac{n-1}{n} A_i^{\text{RLOO}} \quad \Longrightarrow \quad \mathcal{N}(A^{\text{mean}}) = \mathcal{N}(A^{\text{RLOO}}), \tag{12}$$

where $\mathcal{N}$ denotes this ideal normalization with the same active positions and the exact, unregularized two-factor formula.

Corollary B.5 (Equivalence to token-weighted standardization for binary rewards)**.** *Under Proposition 3.1, let μ_{tok} and $\sigma_{\text{tok}} > 0$ be the empirical mean and population standard deviation of the raw advantages over the active tokens of this prompt group. Then the ideal asymmetric coefficients also satisfy $\hat{A}_{i,t} = (A_{i,t} - \mu_{\text{tok}})/\sigma_{\text{tok}}$.*

B.3 The Normalization Mechanism for Actual RLOO Signals

Signed token mass. For binary rewards, the raw positive-to-negative token mass ratio $\bar{T}_+/\bar{T}_-$ reflects unequal class-conditional lengths. The ideal rule balances this mass before importance weighting, masking, and multiplication by score vectors. With one success and nine equally long failures, the $0/1$ RLOO amplitudes are $1, -1/9$ and $\alpha = \beta = 3$, so the negative magnitude increases to $1/3$. Importance weights, masks, and score vectors still affect the update.

Proof of Proposition B.4. For $c > 0$, replacing A by cA preserves the sign sets and token count, multiplies both first sums by c and both squared sums by c^2. Thus κ is unchanged, D becomes $c^2 D$, and the factors become $\alpha/c, \beta/c$, leaving their products with cA unchanged. The baseline identity follows from $r_i - \frac{1}{n}\sum_j r_j = \frac{n-1}{n}(r_i - \frac{\sum_j r_j - r_i}{n-1})$ and $c = (n-1)/n > 0$. □

Proof of Corollary B.5. Write $a = a_+ > 0$, $b = a_- > 0$. The token mean and variance are

$$\mu = \frac{Pa - Nb}{P + N}, \qquad v = \frac{PN(a+b)^2}{(P+N)^2} > 0.$$

The centered coefficients are $N(a+b)/(P+N)$ and $-P(a+b)/(P+N)$. Dividing by $\sqrt{v}$ gives $\sqrt{N/P}$ and $-\sqrt{P/N}$, which match Equation (1). □

From ideal balance to implemented coefficients.

Proposition B.6 (Moment Residuals of the Implemented Normalization)**.** *Suppose both active token-sign groups are nonempty and all inputs are finite. Write*

$$P = \sum_{\mathcal{P}} A_{i,t}, \quad N = -\sum_{\mathcal{N}} A_{i,t}, \quad Q_+ = \sum_{\mathcal{P}} A_{i,t}^2, \quad Q_- = \sum_{\mathcal{N}} A_{i,t}^2,$$

and let $m = |\mathcal{P}| + |\mathcal{N}|$, $\kappa = P/N$, $D = Q_+ + \kappa^2 Q_-$, $\eta = \varepsilon_{\text{num}} > 0$. On the normalization branch, which additionally requires $P, N \geq \eta$ so that the small-sum fallback is inactive, the pre-clamp scales are

$$D_c = Q_+ + \min(\kappa^2 Q_-, C) + \eta, \qquad \alpha_0 = \sqrt{m/D_c}, \qquad \beta_0 = \alpha_0 \kappa,$$

where $C = 10^8$ is the intermediate cap. The advantages scaled by α_0, β_0 have empirical mean zero and second moment D/D_c. In particular, if $\kappa^2 Q_- \leq C$, then the empirical variance v_0 satisfies

$$v_0 = \frac{D}{D+\eta} < 1, \qquad 1 - v_0 = \frac{\eta}{D+\eta} \leq \frac{\eta}{D}. \tag{13}$$

In general,

$$\frac{D}{D_c} - 1 = \frac{\kappa^2 Q_- - \min(\kappa^2 Q_-, C) - \eta}{D_c},$$

so the product cap can change the second moment in either direction relative to one. Let a, b be the final clamped scales and let μ, v_2 be the resulting empirical mean and second moment on the m nonzero active tokens. Then

$$\mu = \frac{aP - bN}{m}, \qquad |\mu| \leq \frac{|a - \alpha_0|P + |b - \beta_0|N}{m},$$
$$v_2 = \frac{a^2 Q_+ + b^2 Q_-}{m}, \qquad \left| v_2 - \frac{D}{D_c} \right| \leq \frac{|a - \alpha_0||a + \alpha_0|Q_+ + |b - \beta_0||b + \beta_0|Q_-}{m}.$$

The empirical variance after factor clamping is $v_2 - \mu^2$. Positive clamping bounds preserve advantage signs and within-sign ordering. These statements describe the normalization branch in exact real arithmetic; when a fallback is taken, the active advantages are unchanged.

Proof. Partition the active index set into positive, negative, and zero advantages. The scaled first and second sums are $aP - bN$ and $a^2 Q_+ + b^2 Q_-$. Substituting $\beta_0 = \alpha_0 P/N$ gives zero first sum, and $\alpha_0^2 = m/D_c$ gives second moment D/D_c. The stabilizer deficit and product-cap identity follow by subtracting one. For the factor-clamp bounds, use

$$aP - bN = (a - \alpha_0)P - (b - \beta_0)N$$

and $a^2 - \alpha_0^2 = (a - \alpha_0)(a + \alpha_0)$, with the analogous identity for b, and apply the triangle inequality. Positivity of the clamped factors gives sign and within-sign order preservation. □

Numerical implementation. On the normalization branch with inactive product and factor clamps, the stabilizer multiplies both ideal amplitudes by $\sqrt{D/(D+\varepsilon_{\text{num}})}$. Their ratio and zero signed mass are preserved, but fixed epsilon breaks exact reward-scale invariance. Proposition B.6 covers clamps and fallbacks.

B.4 Numerical safeguards

The moment formulas and proof are given in Proposition B.6.

The implementation returns the raw signal if there are no active tokens, one sign class is absent, either signed mass is below the numerical epsilon, or a factor is nonfinite. Otherwise it uses epsilon 10^{-8}, product cap 10^{8}, and factor bounds $[10^{-8}, 10]$. The analysis uses exact real arithmetic. A fallback preserves the raw advantages and their token mean, which can be nonzero for unequal lengths.

B.5 Complete update and aggregation

The default algorithm disables uniform scale. For the optional branch, a homogeneous group replaces its signal by R_i/n and skips normalization; an all-zero group contributes zero. The following eligibility condition identifies this branch for binary rewards. Adding a common reward offset changes the replacement signal by that offset divided by n.

Lemma B.7 (Binary uniform-scale eligibility)**.** *For $n \geq 2$ rewards in $\{0,1\}$, the implementation's strict sample-standard-deviation test $\mathrm{std}(r) < \varepsilon_{\text{tol}}$ selects exactly homogeneous groups if $\varepsilon_{\text{tol}} > 0$ and $n\varepsilon_{\text{tol}}^2 \leq 1$. Thus the implemented tolerance 10^{-8} has this property for $2 \leq n \leq 10^{16}$ in exact arithmetic.*

Proof. If k rewards equal one, the sample variance is $k(n-k)/(n(n-1))$. It vanishes for homogeneous groups; for $1 \leq k \leq n-1$, the identity $k(n-k) - (n-1) = (k-1)(n-k-1) \geq 0$ gives $\mathrm{std}(r)^2 \geq 1/n \geq \varepsilon_{\text{tol}}^2$. The strict test therefore excludes mixed groups, including at equality.

□

For the binary update results below, fix a finite response policy p_θ that is differentiable at θ, nonnegative and normalized in a neighborhood of θ. Rewards are fixed independently of the parameter. Write $s(\theta) = \Pr_{p_\theta}(R=1)$ and assume $0 < s < 1$. The response score is $u_\theta(y) = \nabla \log p_\theta(y)$ on positive support and zero on zero-mass responses. Normalization coefficients are held fixed when forming the score update. These are on-policy population statements; the optional uniform branch, when used, selects only homogeneous binary groups.

Proposition B.8 (Complete binary update under class-dependent lengths)**.** *Consider the same finite on-policy model with rewards $R \in \{0,1\}$ and $n = m+1 \geq 2$ iid responses, and divide by this group's total active-token count. Assume that every response with label $q \in \{0,1\}$ has a deterministic active length L_q satisfying $0 < L_q \leq L_{\max}$. Let the positive/negative factors be computed from the length-weighted RLOO advantages using the implemented additive stabilizer, intermediate product cap, factor clamps and identity fallback, with numerical lower clamp $\eta \in (0,1]$ and upper clamp at least η. Let L be the sampled total active-token count and H_i the actual signal before reduction. The response-score coefficient is $C_i = H_i/L$. On-policy, with every prefix mask accepted, the complete expected sequence-score update satisfies*

$$\mathbb{E}\left[\sum_{i=1}^{n} C_i(Y_{1:n}) u_\theta(Y_i)\right] = \Lambda \nabla_\theta s, \qquad \Lambda \geq \frac{\eta}{L_{\max}} > 0,$$

where $s = \Pr_{p_\theta}(R=1)$. This holds both with uniform-scale disabled and with the optional $0/1$ uniform-scale replacement on homogeneous groups. The multiplier may depend on p_θ, the sibling count and the lengths, but it is common to every parameter direction.

The main text gives the conditioning proof. Within-class length variation introduces the coefficient–score covariance in Proposition 3.2.

Corollary B.9 (Local true-reward improvement of the complete binary update)**.** *Fix a prompt, binary rewards $R \in \{0,1\}$, and $n \geq 2$ iid responses from the finite differentiable policy above, with parameters in a real Hilbert space and a policy that is nonnegative and normalized in a neighborhood of θ. Assume $0 < s(\theta) < 1$, where $s(\theta) = \Pr_{p_\theta}(R = 1)$. At the on-policy point, let every prefix mask accept. Allow arbitrary content-dependent active lengths $0 < L(y) \leq L_{\max}$ and retain the actual stabilizer, product cap, factor clamps and identity fallback. The lower factor clamp is $0 < \eta \leq 1$ and the upper clamp is at least η. The optional uniform-scale branch may be enabled or disabled; when enabled, assume its uniformity test selects exactly homogeneous groups (as ensured by Lemma B.7). Using the actual response-score coefficients C_i and variances in Proposition 3.2, set*

$$g := \nabla s(\theta), \qquad U := \mathbb{E}\left[\sum_{i=1}^{n} C_i(Y_{1:n}) u_\theta(Y_i)\right], \qquad f_g(y) := \langle g, u_\theta(y)\rangle.$$

The actual class multiplier satisfies $\Lambda_C \geq \eta/L_{\max} > 0$, and

$$\begin{aligned}\langle g, U\rangle &\geq \Lambda_C \|g\|^2 - \sqrt{V_C\, n\, V_{f,g}} \\ &\geq \frac{\eta}{L_{\max}}\|g\|^2 - \sqrt{V_C\, n\, V_{f,g}}.\end{aligned}$$

If $\sqrt{V_C\, n\, V_{f,g}} < \Lambda_C \|g\|^2$, there exists $r > 0$ such that every $0 < t < r$ satisfies

$$s(\theta + tU) > s(\theta).$$

The simpler condition $\sqrt{V_C\, n\, V_{f,g}} < \eta\|g\|^2/L_{\max}$ is sufficient but can be more conservative than using the actual Λ_C. Under the class-dependent-length premise of Proposition B.8, $g \neq 0$ alone suffices for this local improvement: $U = \Lambda g$ with $\Lambda \geq \eta/L_{\max} > 0$.

Proof. For each fixed sibling draw, the sibling mean reward lies in $[0,1]$. Positive normalization factors are bounded below by η, and the sampled total length is at most $nL_{\max}$. Averaging the resulting success/failure coefficient contrast within each class gives $\Lambda_C \geq \eta/L_{\max}$ even when length varies with content. The optional all-success replacement adds a nonnegative contrast. Apply Proposition 3.2 with $v = g$ and use the finite-policy identity $\mathbb{E}[\mathbf{1}_{\{R=1\}} f_g(Y)] = \|g\|^2$ to obtain the displayed lower bound. A positive right-hand side implies $\frac{d}{dt} s(\theta + tU)\big|_{t=0} = \langle g, U\rangle > 0$. By differentiability, the difference quotient is positive for every sufficiently small positive t, proving strict improvement. For class-dependent lengths, Proposition B.8 directly gives $\langle g, U\rangle = \Lambda\|g\|^2 > 0$ whenever $g \neq 0$.

□

The alignment quantity $\langle g, U\rangle$ is the first-order progress term in the SGA decomposition of Li & Liu (2025b, Section 2). Here conditioning on sibling responses and reward classes identifies that term for the actual normalization and token denominator; Cauchy–Schwarz controls the remaining length–score covariance. Differentiability then gives a step-size neighborhood for the population direction U. The stochastic SGA bound additionally involves the second moment of the entire sampled update. The retained-weight bound below concerns one coefficient, while Theorem 5.1 evaluates finite candidate changes through objective corrections and a population remainder.

Corollary B.10 (Aggregating prompt-specific population updates)**.** *Let μ be a fixed probability law on finitely many prompts. For each prompt x, let s_x be differentiable at θ in a real Hilbert space, $g_x = \nabla s_x(\theta)$, and let its conditional population update have the decomposition $U_x = \Lambda_x g_x + e_x$. The fixed-prompt results above supply Λ_x and the within-class residual e_x for their stated sampling and reduction. Define*

$$J = \sum_x \mu_x s_x, \quad G = \sum_x \mu_x g_x = \nabla J(\theta), \quad U = \sum_x \mu_x U_x, \quad \bar{\Lambda} = \mathbb{E}_\mu \Lambda_x, \quad f_x = \langle G, g_x\rangle.$$

Then

$$\langle G, U\rangle \geq \bar{\Lambda}\|G\|^2 - \sqrt{\mathrm{Var}_\mu(\Lambda_x)\,\mathrm{Var}_\mu(f_x)} - \mathbb{E}_\mu|\langle G, e_x\rangle|. \tag{14}$$

If the right-hand side is strictly positive, then $J(\theta + tU) > J(\theta)$ for every sufficiently small $t > 0$.

Proof. Differentiate the finite sum to identify G. Since $\mathbb{E}_\mu f_x = \|G\|^2$, expanding the pairing gives

$$\langle G, U\rangle = \bar{\Lambda}\|G\|^2 + \mathrm{Cov}_\mu(\Lambda_x, f_x) + \mathbb{E}_\mu\langle G, e_x\rangle.$$

Cauchy–Schwarz bounds the covariance; the last term is bounded below by its negative mean absolute value. A positive directional derivative implies local improvement by differentiability. □

Positive prompt-specific multipliers can reverse the aggregate direction. For example, two equally weighted scalar gradients $1, -2$ and positive multipliers $5, 1$ have $G = -1/2$, $U = 3/2$, and $GU = -3/4$. This algebraic example isolates the effect of reweighting conflicting prompt gradients. A common positive multiplier and $e_x = 0$ give $\langle G, U\rangle = \bar{\Lambda}\|G\|^2 > 0$ when $G \neq 0$.

The preceding mixture averages group-normalized updates. A shared batch token denominator instead couples groups through their lengths. The next result derives the effective coefficient for this global-token reduction.

Proposition B.11 (Effective coefficients for a shared batch denominator)**.** *Sample $K \geq 1$ independent, unmasked prompt groups $B_1, \ldots, B_K$ from the same finite joint law $P(x)\prod_i p_x(y_i)$. Within each group, retain the actual normalized response coefficients $H_i(B)$, including configured safeguards and the binary uniform-scale branch. At the on-policy point with accepted masks, let $F_v(B) = \sum_i H_i(B)\langle v, u_x(Y_i)\rangle$ be the unreduced score contribution in direction v, and let $L(B) > 0$ be the group token count. For independent other groups $B'_1, \ldots, B'_{K-1}$, define*

$$\psi_K(l) := K\,\mathbb{E}\left[\frac{1}{l + \sum_{j=1}^{K-1} L(B'_j)}\right].$$

Then the expected global-token-reduced update satisfies

$$\mathbb{E}\left[\frac{\sum_{k=1}^K F_v(B_k)}{\sum_{k=1}^K L(B_k)}\right] = \mathbb{E}_B[\psi_K(L(B))F_v(B)]. \tag{15}$$

If $0 < l_{\min} \leq L(B) \leq l_{\max}$ for all groups, then $1/l_{\max} \leq \psi_K(L(B)) \leq 1/l_{\min}$. For $K = 1$, $\psi_1(L) = 1/L$. For general K, define $C_i^{(K)}(B) = \psi_K(L(B))H_i(B)$. Using these coefficients in Proposition 3.2 gives the exact prompt-conditional multiplier and residual for the shared-denominator update. Their prompt average can then be evaluated by Corollary B.10 using the coefficients induced by the shared batch denominator.

For binary $0/1$ rewards, suppose both classes have positive probability at the prompt under consideration, with factor-clamp bounds $[\eta, s_{\max}]$ satisfying $0 < \eta \leq \min\{1, s_{\max}\}$. For every $K \geq 1$, the multiplier induced by the actual safeguarded coefficients satisfies

$$\Lambda_x^{(K)} \geq \frac{n\eta}{l_{\max}} > 0. \tag{16}$$

In particular, a uniform per-response length bound $L_{\max}$ permits $l_{\max} = nL_{\max}$ and gives $\Lambda_x^{(K)} \geq \eta/L_{\max}$. The optional homogeneous binary uniform-scale branch preserves this bound; Lemma B.7 gives the condition under which the actual tolerance test selects exactly those groups. For content-dependent lengths, the within-class covariance residual enters the ascent condition alongside this positive multiplier.

Proof. Separate one group from the other $K - 1$ groups in the finite product law. Conditioning on the selected group leaves its numerator fixed. All K coordinates have the same conditional expression, giving Eq. (15). The denominator lies between $Kl_{\min}$ and $Kl_{\max}$, proving the bounds. Expanding F_v gives $C_i^{(K)}$ without assuming independence between length, normalization and score. Conditioning further on the prompt and sibling responses yields the same within-class covariance identity, now with $C_i^{(K)}$. To prove Eq. (16), fix the sibling responses and let their mean reward be $u \in [0, 1]$. The safeguarded factors, including the identity fallback, are at least η. Since $\psi_K \geq 1/l_{\max}$, the current coefficient is at least $\eta(1-u)/l_{\max}$ on success and at most $-\eta u/l_{\max}$ on failure. The uniform-scale branch only adds a nonnegative coefficient on all-success groups. Taking class means gives a contrast of at least $\eta/l_{\max}$ for each coordinate; averaging over siblings and summing the n coordinates proves the bound while retaining the joint dependence of length and score.

□

C REINFORCE PRO: THE IMPLEMENTED PREFIX MASK

C.1 PREFIX STATISTIC AND CAUSAL MASK

Throughout this section the mask uses the current-to-rollout ratio $\rho_t = q(y_t \mid c_t)/p(y_t \mid c_t)$. The same ratio supplies the single differentiable token weight in the objective. Write $\ell_- = 1 - \epsilon_{\text{low}}$ and $\ell_+ = 1 + \epsilon_{\text{high}}$. All prefix counts use active positions. The mask is recomputed at the current policy and detached during differentiation; the token ratio retains its gradient. Rejected positions may become accepted later.

Definition C.1 (Prefix geometric-mean ratio)**.** For current-to-rollout ratios ρ_t, define the prefix statistic

$$\rho_t^{\text{pre}} := \exp\left(\frac{1}{t}\sum_{s=1}^{t}\log\rho_s\right) = \left(\prod_{s=1}^{t}\rho_s\right)^{\frac{1}{t}}, \qquad t \in \{1,\dots,T\}. \tag{17}$$

Here t counts active positions. On a padded horizon, the geometric mean uses $\sum_{s\le t} m_s \log\rho_s$ and denominator $\sum_{s\le t} m_s$; EOS-padding positions have $m_t = 0$. Increasing enumeration of any binary mask's active positions preserves both the prefix average and rejection count. At active positions the count is at least one, so the denominator clamp preserves it. These identities hold for contiguous and noncontiguous masks in exact arithmetic.

Definition C.2 (Causal Prefix Mask)**.** Given thresholds $\epsilon_{\text{low}} \in (0,1)$ and $\epsilon_{\text{high}} > 0$, define the per-position prefix mask

$$M_t^{\text{pre}} := \mathbb{I}\big[\rho_t^{\text{pre}} \in [1-\epsilon_{\text{low}}, 1+\epsilon_{\text{high}}]\big]. \tag{18}$$

Expected log statistic and prefix KL. The conditional-expectation calculation used for sampled KL estimators in Li et al. (2026, Appendix, Sample-Based Estimators) also identifies the population quantity underlying the prefix statistic. Fix a prompt and a nonrandom depth t, with positive mutual support and integrable log ratios. For an autoregressive process with t active positions, let $P_{1:t}^p$ and $P_{1:t}^q$ denote its prefix laws. Then

$$-\mathbb{E}_p[\log\rho_t^{\text{pre}}] = \frac{1}{t}\sum_{s=1}^{t}\mathbb{E}_p\big[D_{\text{KL}}\big(p(\cdot \mid c_s)\|q(\cdot \mid c_s)\big)\big] = \frac{1}{t}D_{\text{KL}}(P_{1:t}^p\|P_{1:t}^q). \tag{19}$$

Indeed, conditioning on c_s gives $\mathbb{E}_p[\log\rho_s \mid c_s] = -D_{\text{KL}}(p(\cdot \mid c_s)\|q(\cdot \mid c_s))$. Summing and using $P_{1:t}^q/P_{1:t}^p = \prod_{s\le t}\rho_s$ proves both equalities. Thus the negative log statistic estimates the per-position prefix KL in expectation. Individual log ratios can cancel, so the two-sided mask tests a sampled statistic rather than bounding every context's KL. In an EOS-padded process the same identity holds with $\rho_s = 1$ after termination and denominator t; the implemented statistic instead divides by its active count. Conditioning on survival or acceptance, or substituting a random active length for t, changes the expectation and requires that conditioning or denominator to remain explicit.

C.2 RETAINED-WEIGHT BOUNDS

Proposition C.3 (Unbounded retained weights under prefix acceptance)**.** *Fix any finite mask thresholds $0 < \ell_- \le 1 \le \ell_+$. Over normalized two-action, two-step policies with strictly positive support, the retained coefficient $M_2\rho_2$ and its rollout second moment are unbounded under prefix acceptance. Specifically, for every $K \ge 1$ there are policies $p = \pi_{\text{roll}}$ and q such that on $y = (1,1)$,*

$$\rho_1 = K^{-1}, \quad \rho_2 = K, \quad M_2 = 1, \quad p(y) = \frac{K}{(K+1)^2}, \qquad \mathbb{E}_p[(M_2\rho_2)^2] \ge \frac{K^3}{(K+1)^2} \ge \frac{K}{4}.$$

Proof. At the empty history set $p(1) = K/(K+1)$ and $q(1) = 1/(K+1)$; at every nonempty history interchange these probabilities. Assign the complementary probability to action 0. Both policies are normalized and strictly positive. On $(1,1)$ the ratio product is one, so the second prefix is accepted at every such threshold. Its contribution to the second moment is $p(1,1)K^2 = K^3/(K+1)^2$. All other contributions are nonnegative, and $(K+1)^2 \le 4K^2$ for $K \ge 1$. Both the coefficient and its second moment grow without bound as K increases. For this same pair, the root token-level TV is $(K-1)/(K+1)$, which approaches one while the displayed second prefix remains accepted. □

As in the token/sequence IS analysis of Li & Liu (2025c, Analyses 1–2), we use conditional ratio normalization to control second moments. The Seq-MIS analysis of Li & Liu (2025a, Section 1.3) squares a pointwise weight envelope. For the retained coefficient alone, we instead use $M_t\rho_t^2 \le CM_t\rho_t$ on histories with envelope C, keeping the accepted importance mass when averaging. Prefix reentry contributes a separate term. Multiplying by the RLOO signal and policy score introduces the additional factors of a gradient estimator.

Theorem C.4 (Retained-weight moments and lower-side reentry)**.** *Consider the finite normalized policy pair $(p, q) = (\pi_{\mathrm{roll}}, \pi_\theta)$ with mutual support. Let $\rho_t = q(y_t \mid y_{<t})/p(y_t \mid y_{<t})$, $W_t = \prod_{s\le t}\rho_s$, and $M_t = \mathbf{1}\{-\tau_- \le t^{-1}\log W_t \le \tau_+\}$ for $\tau_\pm \ge 0$. Set $W_0 = M_0 = 1$ and define*

$$C_t = \exp\{t\tau_+ + (t-1)\tau_-\}, \qquad \mathcal{R}_t^- = \mathbb{E}_p\big[\mathbf{1}\{\log W_{t-1} < -(t-1)\tau_-\}M_t\rho_t^2\big].$$

If $\log W_{t-1} \ge -(t-1)\tau_-$ and $M_t = 1$, then $\rho_t \le C_t$. Moreover,

$$\mathbb{E}_p[(M_t\rho_t)^2] \le C_t\mathbb{E}_p[M_t\rho_t] + \mathcal{R}_t^- \le C_t + \mathcal{R}_t^-. \tag{20}$$

If $t \le T$, $T\tau_+ \le a$, $T\tau_- \le b$, and $\mathcal{R}_t^- \le r$, then $\mathbb{E}_p[(M_t\rho_t)^2] \le e^{a+b} + r$. If no prefix below the lower threshold is followed by an accepted token on rollout support, $\mathcal{R}_t^- = 0$; upper-side reentry is allowed.

Proof. Step 1: local envelope. For an accepted current token, subtracting the previous log-product from $\log W_t \le t\tau_+$ gives $\rho_t \le C_t$ whenever the preceding prefix is above its lower threshold. *Step 2: conditional normalization.* The local envelope bounds the conditional second moment by $C_t\mathbb{E}_p[M_t\rho_t \mid h]$. At every history, $\mathbb{E}_p[M_t\rho_t \mid h] \le \sum_y q(y \mid h) = 1$. *Step 3: history decomposition.* Average the first term over histories above the lower threshold and enlarge it to all histories by nonnegativity; the complementary histories contribute exactly $\mathcal{R}_t^-$. This gives the retained-mass bound, and averaging the unit first-moment bound gives its coarser form. The horizon bound follows from $t\tau_+ + (t-1)\tau_- \le T\tau_+ + T\tau_- \le a + b$. An upper excursion satisfies $\log W_{t-1} > (t-1)\tau_+$, so an accepted next token has $\log\rho_t < \tau_+$ and requires no additional moment term. For EOS-padded responses, M_t in this calculation is the full-tree prefix mask. At active positions it agrees with the implemented mask, so the active coefficient is $m_tM_t\rho_t$. Its unconditional second moment is at most $\mathbb{E}_p[(M_t\rho_t)^2]$, since $m_t \in \{0, 1\}$. The bound therefore also controls the implemented active coefficient; its reentry term is computed on the full padded tree, including any post-EOS transitions. □

Example C.5 (Nonzero retention with bounded moments)**.** Fix a horizon $T \ge 1$, symmetric log-threshold $\tau \ge 0$, and $K > 0$ with $\log K > T\tau$. On an alphabet $\{0, 1, 2\}$, set the first-token probabilities to

$$p_1 = \left(\frac{1}{2}, \frac{1}{2(K+1)}, \frac{K}{2(K+1)}\right), \qquad q_1 = \left(\frac{1}{2}, \frac{K}{2(K+1)}, \frac{1}{2(K+1)}\right).$$

At every nonempty history, let both policies be uniform on the three tokens. These policies are normalized and strictly positive. Their first-token ratios are $(1, K, K^{-1})$; every later token ratio is one. Hence, at every $1 \le t \le T$, the prefix mask accepts exactly the paths whose first token is zero. In particular,

$$\Pr_p(M_t = 1) = \frac{1}{2}, \qquad \mathcal{R}_t^- = 0, \qquad \mathbb{E}_p[(M_t\rho_t)^2] = \frac{1}{2}. \tag{21}$$

The unmasked first-token second moment is $\mathbb{E}_p[\rho_1^2] = (K + K^{-1})/2$, which can be arbitrarily large.

Proof. Along a path starting with token 1 or 2, the cumulative log-ratio is respectively $\log K$ or $-\log K$ at every depth. Its average lies outside $[-\tau, \tau]$ through depth T. Along the token-0 branch it is zero. Acceptance therefore depends only on the first token and never changes within this horizon. The neutral branch has rollout mass $1/2$, and its retained weights are all one, proving Eq. (21). Summing $p_1(a)(q_1(a)/p_1(a))^2$ gives the unmasked second moment. □

C.3 Consequences for causal rejection

The prefix log-average determines how acceptance evolves along a response.

Lemma C.6 (Stability of Prefix Geometric Means)**.** *Fix thresholds $\epsilon_{\text{low}} \in (0,1)$ and $\epsilon_{\text{high}} > 0$, and define ρ_t^{pre} as in Eq. (17). If*

$$\rho_t^{\text{pre}} \in [1-\epsilon_{\text{low}}, 1+\epsilon_{\text{high}}] \qquad \textit{and} \qquad \rho_{t+1} \in [1-\epsilon_{\text{low}}, 1+\epsilon_{\text{high}}],$$

then $\rho_{t+1}^{\text{pre}} \in [1-\epsilon_{\text{low}}, 1+\epsilon_{\text{high}}]$.

Proof. The identity

$$\log \rho_{t+1}^{\text{pre}} = \frac{t}{t+1} \log \rho_t^{\text{pre}} + \frac{1}{t+1} \log \rho_{t+1}$$

is a convex combination of two numbers in $[\log(1-\epsilon_{\text{low}}), \log(1+\epsilon_{\text{high}})]$. Exponentiating preserves the bounds. □

The same cumulative log-ratio gives the following early- and late-deviation bounds.

Theorem C.7 (Conditional prefix rejection patterns)**.** *Compare the prefix mask M_t^{pre} of Definition C.2, the token-level IcePop mask M_t^{tok}, and the sequence mask*

$$M_t^{\text{seq}} = \mathbf{1}\left\{ \exp\left(T^{-1} \sum_{s=1}^{T} \log \rho_s \right) \in [1-\epsilon_{\text{low}}, 1+\epsilon_{\text{high}}] \right\}.$$

All use the same current-to-rollout ratios and interval, with $\epsilon_{\text{low}} \in (0,1)$, $\epsilon_{\text{high}} > 0$. Assume mutual rollout/current support and set $\tau_+ = \log(1+\epsilon_{\text{high}})$, $\tau_- = -\log(1-\epsilon_{\text{low}})$.

(a) ***Early deviation.*** *Suppose ρ_1 is outside the interval and $|\log \rho_t| \le \varepsilon$ for $t > 1$, where $0 \le \varepsilon \le \min(\tau_+, \tau_-)$. IcePop rejects only position one; the prefix mask also rejects it. If $\log \rho_1 > \tau_+$, prefix rejection persists whenever*

$$t < \frac{\log \rho_1 + \varepsilon}{\tau_+ + \varepsilon}. \tag{22}$$

If $\log \rho_1 < -\tau_-$, it persists whenever

$$t < \frac{-\log \rho_1 + \varepsilon}{\tau_- + \varepsilon}. \tag{23}$$

Thus the prefix mask rejects at least as many tokens as IcePop in this regime.

(b) ***Late deviation.*** *Suppose ρ_{t^*} is outside the interval for some $t^* \in \{1, \dots, T\}$, while $|\log \rho_t| \le \varepsilon$ for $t \ne t^*$, where $0 \le \varepsilon \le \min(\tau_+, \tau_-)$. Every prefix before t^* is accepted, so*

$$|\{t : M_t^{\text{pre}} = 0\}| \le T - t^* + 1, \qquad |\{t : M_t^{\text{seq}} = 0\}| \in \{0, T\}.$$

Proof. Write $z_s = \log \rho_s$. The three masks test, respectively, z_t, $t^{-1}\sum_{s\le t} z_s$, and $T^{-1}\sum_{s\le T} z_s$ against the same interval $[-\tau_-, \tau_+]$.

(a) Early deviation. Every $t > 1$ is accepted by the token mask because $z_t \in [-\varepsilon, \varepsilon] \subseteq [-\tau_-, \tau_+]$; position one is rejected by both token and prefix masks. For every $1 \le t \le T$, the bounded later increments give

$$\frac{z_1 - (t-1)\varepsilon}{t} \le \frac{1}{t}\sum_{s\le t} z_s \le \frac{z_1 + (t-1)\varepsilon}{t}.$$

The left bound exceeds τ_+ under Equation (22); the right bound is below $-\tau_-$ under Equation (23). Both denominators are positive. The strict windows give persistent rejection; rejection at position one gives the token-count comparison.

(b) Late deviation. For $t < t^*$ every summand of the prefix average lies in the acceptance interval, so its average does too. The rejected positions are therefore a subset of $\{t^*, \dots, T\}$, of size at most $T - t^* + 1$. The constant sequence mask rejects zero or T tokens. □

Response length and retained token mass. Li et al. (2026, Appendix, Length-Neutral Trust Region Masking) distinguish a fixed threshold from length-independent acceptance. For Pro, define the retained fraction of a response with ℓ active tokens by $r_\ell = \ell^{-1}\sum_{t=1}^{\ell} M_t$. For random active length L, linearity gives the exact conditional identity

$$\mathbb{E}_p[r_L \mid L = \ell] = \frac{1}{\ell}\sum_{t=1}^{\ell} \Pr_p(M_t = 1 \mid L = \ell).$$

No independence between prefix decisions is needed. Under the late-deviation premises of Theorem C.7, $r_\ell \geq (t^*-1)/\ell$. Under its early upper-deviation premises, writing $b = (\log\rho_1+\varepsilon)/(\tau_+ +\varepsilon)$ gives $r_\ell \leq 1-\ell^{-1}|\{t \in \{1,\dots,\ell\} : t < b\}|$; the lower-deviation case uses $b = (-\log\rho_1+\varepsilon)/(\tau_- + \varepsilon)$. These follow by counting the accepted prefix or the guaranteed rejection window. They describe how deviation location and duration affect retained mass. For a whole-sequence maximum rule, independent per-token violations with probability v give acceptance $(1-v)^\ell$, as in TRM's illustrative model. Pro uses dependent running averages with possible reentry, so its retention is governed by the displayed conditional average instead.

The interaction with Max can be stated exactly on a sampled prompt group. Let $\mathcal{I}_+$ and $\mathcal{I}_-$ be its positive and negative active-token sets, $S_\pm = \sum_{k\in\mathcal{I}_\pm} |A_k|$, and assume both are nonempty. For the ideal coefficients, write $B = \alpha S_+ = \beta S_- > 0$ and define $\eta_\pm = S_\pm^{-1}\sum_{k\in\mathcal{I}_\pm} M_k|A_k|$. Then, before importance weighting and multiplication by score vectors,

$$\sum_k M_k\hat{A}_k = B(\eta_+ - \eta_-), \qquad \left|\sum_k M_k\hat{A}_k\right| \leq B\big[(1-\eta_+) + (1-\eta_-)\big]. \tag{24}$$

Substitute $\hat{A}_k = \alpha A_k$ or βA_k on the two sign sets to obtain the equality; the inequality follows by the triangle inequality. For binary rewards, advantages are constant within each sign class, so $\eta_\pm$ are precisely its retained token fractions. Masking preserves ideal signed-mass balance exactly when those fractions agree. For actual protected coefficients, the corresponding equality is $\sum_k M_k\hat{A}_k = \alpha S_+\eta_+ - \beta S_-\eta_-$. This identifies the effect of unequal class retention without altering the algorithm. The reward bound uses the importance-weighted discarded magnitude $\sum_k \rho_k(1-M_k)|\hat{A}_k|$ in LD_λ, rather than an unweighted retention fraction, to account for its objective cost.

D COMBINED UPDATE AND REWARD-IMPROVEMENT CONDITION

Following the surrogate-minus-remainder structure of Qi et al. (2026); Li et al. (2026), we express the masked RLOO objective through its raw surrogate, retaining normalization and reduction effects. The masking term adapts the discarded-contribution analysis of Li & Liu (2025a, Section 1.3) to tokenwise masks.

For each candidate q, recompute $M_k(q)$ using the same ratio $\rho_k(q) = q/p$ that appears in the objective:

$$\mathcal{L}^{\mathrm{ProMax}}(q) = -\sum_k \omega_k M_k(q)\rho_k(q)\hat{A}_k. \tag{25}$$

The sampled advantages and nonnegative reduction weights are fixed. The implementation detaches M and $\hat{A}$, and differentiates ρ. Consequently the update is $\sum_k \omega_k M_k\rho_k\hat{A}_k\nabla\log q$. This agrees with the ordinary objective gradient within any open region where every sampled mask decision is constant. No derivative of a hard threshold is required to define the implemented update at its boundary.

Proposition D.1 (Finite-Batch ProMax Objective Decomposition)**.** *Let $k = (i,t)$ index a sampled response and position. Write $A_k = R_i - b_i$ for its raw RLOO signal, $\hat{A}_k$ for the actual scaled signal, $\rho_k = q/p \geq 0$, and $M_k = M_k(q) \in \{0,1\}$. Let $\omega_k \geq 0$ include the action mask and reduction weights; for the active-token mean, $\omega_{i,t} = m_{i,t}/\sum_{j,s} m_{j,s}$. Define*

$$S_{\mathrm{PM}} = \sum_k \omega_k M_k\rho_k\hat{A}_k, \qquad U = \sum_k \omega_k\rho_k A_k,$$
$$G = \sum_k \omega_k\rho_k(1-M_k)A_k, \qquad N = \sum_k \omega_k\rho_k M_k(\hat{A}_k - A_k).$$

Then

$$S_{\mathrm{PM}} = U - G + N, \tag{26}$$

and

$$|S_{\mathrm{PM}} - U| \le \sum_k \omega_k \rho_k (1 - M_k)|A_k| + \sum_k \omega_k \rho_k M_k |\hat{A}_k - A_k|.$$

Proof. Add and subtract $\sum_k \omega_k \rho_k M_k A_k$. The triangle inequality and nonnegative weights give the bound. The calculation is pointwise, so the mask may depend on q, the prefix, and the sampled action. □

D.1 Scaling, masking, and the combined reward bound

The proof of Theorem 5.1 connects the DPPO/TRM raw surrogate to the actual objective in three steps: a pointwise residual, conditional cancellation of each independent sibling baseline, and the sequence-error bound. The deterministic reference scale $\lambda > 0$ stays outside the expectation; sampled Max factors, current-policy masks and token counts stay inside D_λ. Common scaling $H = \lambda A$ removes the scaling residual, and full retention removes the rejection residual. The following support and reduction conventions specify that correspondence for variable-length responses and accumulated microbatches.

D.2 Support and EOS in the population bound

Set $m(h) = \mathbf{1}\{\mathrm{EOS} \notin h\}$, including the EOS-emitting action. On sampled support, mutual support gives $\rho = q/p$; at the rollout reference $q = p$, this ratio and its prefix geometric mean equal one. After EOS, the common absorbing transition has probability one, so $\rho = 1$ and $m(h)(\rho - 1) = \rho - 1$ for every raw RLOO signal. Batches containing a zero-mass response have zero sampling mass and contribute zero under any finite ratio convention. For a nonempty horizon, the first response token is active, giving active length at least one.

Active-token reduction. For a group contribution Z and active token count L, averaging group token means gives $\mathbb{E}[Z/L]$, while a population token mean is $\mathbb{E}[Z]/\mathbb{E}[L]$. Their difference is

$$\mathbb{E}[Z/L] - \frac{\mathbb{E}[Z]}{\mathbb{E}[L]} = -\frac{\mathbb{E}[(L - \mathbb{E}[L])Z/L]}{\mathbb{E}[L]}.$$

The population theorem retains the random L inside its expectation; Proposition B.11 treats the shared-batch expected update.

For microbatch j, let $N_j > 0$ be its response count, $L_j > 0$ its active-token count, and Z_j its unreduced surrogate contribution at the same candidate policy. Write $N_\Sigma = \sum_j N_j$ and $L_\Sigma = \sum_j L_j$. Token-count weighting reconstructs the full-batch objective:

$$\sum_j \frac{L_j}{L_\Sigma} \frac{Z_j}{L_j} = \frac{\sum_j Z_j}{L_\Sigma}.$$

The accompanying implementation's dynamic batching instead assigns response-count weights N_j/N_Σ to its microbatch token means. The resulting difference is exactly

$$\sum_j \frac{N_j}{N_\Sigma} \frac{Z_j}{L_j} - \frac{\sum_j Z_j}{L_\Sigma} = \sum_j \left(\frac{N_j}{N_\Sigma} - \frac{L_j}{L_\Sigma} \right) \frac{Z_j}{L_j}.$$

These identities follow by distributing the sums. The finite-batch decomposition uses weights $m_{i,t}/L_\Sigma$ or $N_j m_{i,t}/(N_\Sigma L_j)$, respectively. With equal response counts per data-parallel rank, j can index all rank–microbatch pairs. Denote the displayed difference by $E_{\mathrm{red}}(q)$ and hold the sampled partition fixed. Substituting $\Delta S_{\mathrm{batch}} = \Delta S_{\mathrm{dynamic}} - [E_{\mathrm{red}}(q) - E_{\mathrm{red}}(p)]$ into Theorem 5.1, with token-weighted $D_{\lambda,\mathrm{batch}}$, gives the reward bound for the dynamic objective.