V-TRACE POLICY OPTIMIZATION: WHEN MULTI-STEP CORRECTION HELPS, AND WHY

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning for language models increasingly departs from the on-policy ideal: rollouts are reused across optimizer passes, actors lag learners, and inference and training engines disagree on token probabilities. PPO and GRPO correct only the current token, whereas exact trajectory importance sampling corrects the whole sequence at prohibitive variance. We show that prefix, current-action, and suffix ratios reweight the visited history, the sampled action, and its continuation, respectively. We introduce *V-Trace Policy Optimization* (V-TRACE PO), which attaches a coefficient-truncated suffix trace directly to the per-token score. One reverse scan computes the trace, and on complete terminal-reward responses it requires no learned critic; an exact identity says which continuation policy the truncated trace targets, and whatever the truncation, the correction adds no bias to the per-token score.

We evaluate it in two families of individually preregistered experiments: one in which the same rollouts are reused across several optimizer passes, and one in which every sample is fresh but the behavior policy differs from the learner by construction—a lagging generator, a hotter sampler, or sequential minibatches—together with a sweep of the truncation cap, so the correction is also tested at full strength. The change-of-measure argument that motivates the estimator applies to the second family and not the first. The results come out the other way round: V-TRACE PO helps under reuse, and reliably in none of our fresh-sample controls, where widening the correction is if anything harmful.

We analyze both halves. On fresh samples the correction does not pay: at every truncation we test, the variance it adds exceeds the bias it removes — even at the mildest mismatch we can construct, the numerical disagreement between the inference and training engines. Under reuse it helps for a reason the construction does not predict: the trace silently becomes an asymmetric learning rule, amplifying whatever an earlier pass over the same batch already did — promoting what succeeded and damping what failed. Under heavy sample reuse the improvement of V-TRACE PO over GRPO is substantial, and it comes with a wider usable trust region — hence less clipping — where that same widening degrades GRPO.

1 INTRODUCTION

Because verifiable reasoning rollouts are costly, systems reuse them across optimizer passes or sequential minibatches, replay them from a buffer, or decouple actors from learners (Schmitt et al., 2020; Arnal et al., 2026; Sheng et al., 2025; Fu et al., 2025), creating behavior–learner mismatch. Even identical checkpoints can assign different token probabilities across inference and training engines (Yao et al., 2025; He & Thinking Machines Lab, 2025). Full-trajectory importance sampling is exact but its variance grows with sequence length (Metelli et al., 2020; Huang & Jiang, 2020), whereas Proximal Policy Optimization (PPO) and Group Relative Policy Optimization (GRPO) use current-token ratios (Schulman et al., 2017; Shao et al., 2024; Ahmadian et al., 2024). In autoregressive policies, prefix, current and suffix ratios respectively reweight the visited history, the sampled action, and its continuation.

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.

V-TRACE PO at a glance. For a response sampled by behavior policy μ, define the current-token ratio and adaptive continuation coefficient

$$\rho_t = \frac{\pi_\theta(a_t \mid x_t)}{\mu(a_t \mid x_t)}, \qquad c_t = \min\{\bar{c}, \rho_t\}.$$

Compute the suffix trace by one backward scan and detach its ratio to the current token:

$$\widetilde{\rho}_{T-1} = \rho_{T-1}, \qquad \widetilde{\rho}_t = \rho_t + c_t(\widetilde{\rho}_{t+1} - 1), \qquad w_t = \mathrm{sg}\left(\frac{\widetilde{\rho}_t}{\rho_t}\right). \tag{1}$$

For detached response advantage A_b, valid-token mask M_{bt} and $Z = \sum_{b,t} M_{bt}$, the clipped objective is

$$\mathcal{J} = \frac{1}{Z} \sum_{b,t} M_{bt}\, w_{bt} \min\{\rho_{bt} A_b, \mathrm{clip}(\rho_{bt}, 1-\epsilon, 1+\epsilon) A_b\}. \tag{2}$$

The trust region acts on ρ_{bt} while the detached correction rides outside it, so the score coefficient is $A_b \widetilde{\rho}_{bt}$ before clipping. Here $\bar{c} = 0$ recovers current-token PPO/GRPO and $\bar{c} = \infty$ the full suffix ratio $\widetilde{\rho}_t = \prod_{s=t}^{T-1} \rho_s$; we default to $\bar{c} = 1$ and sweep the cap in Section 6.2.

Section 2 gives the estimator's switching interpretation and its exact V-trace–Retrace links (Espeholt et al., 2018; Munos et al., 2016).

Empirically, suffix correction pays under *sample reuse*: taking E optimizer passes over one sampled batch instead of a single pass. At $E = 8$, $\bar{c} = 1$ and a wide trust region, a matched GRPO baseline gives back much of the accuracy it reached at shallower reuse while V-TRACE PO keeps improving, and the gap grows monotonically with E at both widths we test.

The more consequential finding is that the benefit has a different source than its derivation suggests. The change-of-measure argument predicts a gain whenever Q_t^μ and Q_t^π differ materially, regardless of reuse. Three studies at $E = 1$ construct that situation—generator staleness over two orders of magnitude, a sampler run hotter than the policy, and sequential minibatches—and a fourth sweeps the truncation cap at fixed mismatch; none yields a reliable gain. Section 4 identifies why. When all samples are fresh, the variance the trace adds outweighs the bias it removes. Under reuse, an earlier pass has already pushed each response's tokens toward its own advantage, so its ratio deviations share that sign and accumulate in the trace, weighting updates asymmetrically in the sign of the advantage. We measure that asymmetry and find it confined to the reuse regime.

Contributions. *An anatomy of the ratio.* Four exact finite-horizon identities (Prop. 2.1) separate the three things a product of ratios can correct — the prefix-state law, the current action, and the continuation — and isolate the suffix product as the one that corrects the continuation while leaving the state law under μ. *A trace that carries it.* We attach a coefficient-truncated suffix recurrence to the score (Eq. 7), computed in a single reverse scan, with c_t bounded by $0 \le c_t \le \rho_t$ so that it interpolates between the local ratio and the full suffix product. It adds no bias to the score for any coefficient rule (Prop. 2.2), and it recovers V-trace and Retrace as choices of c_t, with a switching interpretation that needs no critic on terminal-reward responses (§2). The resulting optimizer adds no rollout cost to a GRPO trainer. *What actually drives the benefit.* We identify and measure the asymmetry above, and show that it rather than the change of measure is what tracks the empirical gain (§4).

2 RATIO ANATOMY AND SUFFIX TRACES

Consider an undiscounted finite response $(x_0, a_0, \ldots, x_{T-1}, a_{T-1})$ with terminal reward R; equivalently, $\gamma = 1$. Appendix J.3 generalizes the construction to dense rewards, where each reward carries only the trace accumulated up to its own arrival. The behavior policy is μ, the learner is π_θ, and

$$\rho_t = \frac{\pi_\theta(a_t \mid x_t)}{\mu(a_t \mid x_t)}, \qquad s_t = \nabla_\theta \log \pi_\theta(a_t \mid x_t).$$

We write $\pi = \pi_\theta$ below. Write $\rho_{i:j} := \prod_{k=i}^{j} \rho_k$, with an empty product equal to one. We assume target support is covered by μ on behavior-reachable histories. In an LLM context, x_t contains the prompt and all tokens preceding position t. For a policy ν, d_t^ν is the distribution of these prefix states induced by the prompt distribution and actions $a_0, \ldots, a_{t-1} \sim \nu$; for $\nu = \mu$ it is the state distribution at which the token-t score is evaluated. Let $Q_t^\nu(x, a)$ be the expected terminal return after fixing $(x_t, a_t) = (x, a)$ and following ν thereafter.

Proposition 2.1 (Three roles of importance ratios) *For every $t \in \{0, \ldots, T-1\}$,*

$$\mathbb{E}_\mu[R\,\rho_{0:T-1} s_t] = \mathbb{E}_{x_t \sim d_t^\pi,\, a_t \sim \pi}\big[Q_t^\pi(x_t, a_t)\, s_t\big], \tag{3}$$

$$\mathbb{E}_\mu[R\,\rho_{t:T-1} s_t] = \mathbb{E}_{x_t \sim d_t^\mu,\, a_t \sim \pi}\big[Q_t^\pi(x_t, a_t)\, s_t\big], \tag{4}$$

$$\mathbb{E}_\mu[R\,\rho_{0:t} s_t] = \mathbb{E}_{x_t \sim d_t^\pi,\, a_t \sim \pi}\big[Q_t^\mu(x_t, a_t)\, s_t\big], \tag{5}$$

$$\mathbb{E}_\mu[R\,\rho_t s_t] = \mathbb{E}_{x_t \sim d_t^\mu,\, a_t \sim \pi}\big[Q_t^\mu(x_t, a_t)\, s_t\big]. \tag{6}$$

Color marks the law each factor carries: blue for the target (learner) law, red for the behavior law. The current action is drawn from π in every case; a weight changes the prefix-state law (Eq. (5)), the continuation (Eq. (4)), both (Eq. (3)), or neither (Eq. (6)). Proofs and support qualifications are in Appendix F.

The local ratio is therefore an exact gradient for a behavior-anchored surrogate, but it leaves both the prefix-state law and future continuation under μ. Full-trajectory correction changes all three at the cost of the longest product of ratios, and so the highest variance. We retain d_t^μ, correct the current action, and control how far the future continuation is corrected. Prefix correction is outside this estimator by construction: it reweights the states at which the policy is improved rather than the improvement made at each state, so under the coverage assumption above and exact realizability it alters the optimization path but not which policy is optimal. Because every arm in our experiments retains the same d_t^μ, the prefix factor is common to all reported comparisons and cannot account for any difference between them.

For fixed μ and π, let each trace coefficient be a state-action function $c_t = c_t(x_t, a_t)$ satisfying $0 \le c_t \le \rho_t$ on behavior support. Define the cumulative suffix trace

$$\widetilde{\rho}_t = \rho_t + \sum_{s=t+1}^{T-1} \left(\prod_{j=t}^{s-1} c_j \right) (\rho_s - 1), \qquad \widetilde{\rho}_{T-1} = \rho_{T-1}, \quad \widetilde{\rho}_t = \rho_t + c_t(\widetilde{\rho}_{t+1} - 1) \quad (t < T-1). \tag{7}$$

The second equality gives a single reverse scan. The associated unclipped sample direction is

$$\widehat{G}_c = R \sum_{t=0}^{T-1} \widetilde{\rho}_t s_t. \tag{8}$$

For the adaptive family $c_t = \min\{\bar{c}, \rho_t\}$ and any $\bar{c} \ge 0$, $|\widetilde{\rho}_t - \rho_t| \le \sum_{s=t+1}^{T-1} \bar{c}^{\,s-t} |\rho_s - 1|$, bounding the deviation in both directions since $\rho_s - 1$ carries its own sign (Prop. B.3). Thus a unit cap (i.e. $\bar{c} = 1$) makes deviations at distant s additive rather than multiplicative; it moderates but does not uniformly bound the trace.

Proposition 2.2 (Trace normalization and score centering) *Let $\mathcal{F}_t$ be the prompt and history immediately before a_t is sampled. Then for every decision time*

$$\mathbb{E}_\mu[\widetilde{\rho}_t \mid \mathcal{F}_t] = 1, \qquad \mathbb{E}_\mu[\widetilde{\rho}_t s_t \mid \mathcal{F}_t] = 0.$$

Both hold for any integrable adapted c_t; the bound $0 \le c_t \le \rho_t$ matters only for the switching interpretation below.

The second identity is not immediate: $\mathbb{E}_\mu[\rho_t f(x_t, a_t) s_t]$ is nonzero for general f, and $\widetilde{\rho}_t$ does depend on a_t through the continuation — but that dependence only multiplies $\widetilde{\rho}_{t+1} - 1$, which the first identity centers at $\mathcal{F}_{t+1}$. Because this holds for every coefficient rule, a measured mean trace of one is no evidence the correction is doing anything; only its covariation with the advantage is (Section 4).

To identify the continuation an intermediate trace induces, split each coefficient into a change of measure on the current action and a factor controlling the continuation. Set $\lambda_t(x, a) = c_t(x, a)/\rho_t(x, a)$ when $\rho_t(x, a) > 0$, and set $\lambda_t(x, a) = 0$ otherwise. Then $\lambda_t \in [0, 1]$, $c_t = \rho_t \lambda_t$, and

$$\widetilde{\rho}_t = \rho_t\big[(1 - \lambda_t) + \lambda_t \widetilde{\rho}_{t+1}\big], \qquad t < T - 1.$$

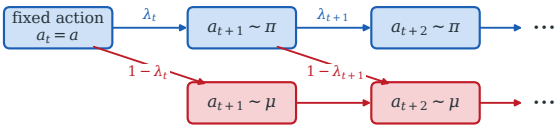

Figure 1: **The continuation a trace coefficient induces.** $Q_t^c(x,a)$ is the value of taking action a in state x at time t and then continuing under π with probability $\lambda_t = c_t/\rho_t$, or switching permanently to μ with probability $1-\lambda_t$; the same decision is repeated at $t+1, t+2, \ldots$ after each further π-action.

Thus ρ_t changes the current action law from μ to π, whereas λ_t controls whether that correction continues beyond the current action.

Specifically, after the fixed current action a_t, continue with π for the next action with probability λ_t; otherwise switch permanently to μ. Repeat this decision after each subsequent π-action, with no decision after the terminal action. Define $Q_t^c(x,a)$ as the expected terminal return under this history-dependent switching continuation. The superscript records the trace rule, not a stationary policy, in direct analogy with the continuation-policy superscripts in Q_t^μ and Q_t^π. Figure 1 shows the timing; this switching process is an interpretation and is never simulated by the algorithm, and Appendix B gives the equivalent recursion.

Theorem 2.3 (Finite-horizon switching target) *For every behavior-supported* (x,a),

$$\mathbb{E}_\mu[R\widetilde{\rho}_t \mid x_t = x, a_t = a] = \rho_t(x,a) Q_t^c(x,a), \quad \mathbb{E}_\mu[\widehat{G}_c] = \sum_t \mathbb{E}_{\substack{x_t \sim d_t^\mu \\ a_t \sim \pi}}[Q_t^c(x_t,a_t) s_t]. \tag{9}$$

For the adaptive rule $c_t = \min\{\bar{c}, \rho_t\}$, $\lambda_t = \min\{1, \bar{c}/\rho_t\}$ when $\rho_t > 0$: $\bar{c} = 0$ switches immediately and gives $Q_t^c = Q_t^\mu$, the formal $\bar{c} = \infty$ endpoint never switches along target-supported continuations and gives $Q_t^c = Q_t^\pi$ there, and intermediate c mixes the two laws. The ordering $Q_t^\mu \le Q_t^c \le Q_t^\pi$ follows once the one-step target-policy backup improves on the behavior action value, but not from the interpolation alone (Theorem I.3). Equation (9) is a semigradient: it differentiates the current-action probability while holding the continuation target and the state distribution fixed, so it is a surrogate anchored at μ rather than an on-policy return gradient.

The same trace defines an action-conditioned extension of IMPALA's state-value V-trace construction (Espeholt et al., 2018), written here as a finite sum rather than a recursion. Write $c_{i:j} := \prod_{k=i}^{j} c_k$, again with empty products equal to one, and let $V = (V_0, \ldots, V_T)$ be time-indexed state values, with the terminal value carrying the return, $V_T := R$. We introduce the *state-action V-trace operator* $\mathcal{V}_c^{\mathrm{sa}}$, which maps V to the unnormalized, action-conditioned target

$$[\mathcal{V}_c^{\mathrm{sa}} V]_t(x,a) = V_t(x) + \mathbb{E}_\mu\left[\sum_{s=t}^{T-1} c_{t:s-1}\big(\rho_s V_{s+1}(x_{s+1}) - V_s(x_s)\big) \,\middle|\, x_t = x, a_t = a\right].$$

Two endpoints matter. At $c = 0$ it reduces to $[\mathcal{V}_{c=0}^{\mathrm{sa}} V]_t = \rho_t [\mathcal{B}V]_t$, where $[\mathcal{B}V]_t(x,a) := \mathbb{E}[V_{t+1}(x_{t+1}) \mid x_t = x, a_t = a]$ lifts a state value to an action value; at full coefficients $c = \rho$ it gives $[\mathcal{V}_{c=\rho}^{\mathrm{sa}} V]_t = \rho_t Q_t^\pi$, independently of V. At the behavior value the sampled return realizes the operator outright,

$$[\mathcal{V}_c^{\mathrm{sa}} V^\mu]_t(x,a) = \mathbb{E}_\mu[R\widetilde{\rho}_t \mid x_t = x, a_t = a],$$

its Monte-Carlo realization, with the sampled return in place of each behavior value, telescoping pathwise to $R\widetilde{\rho}_t$, which is why Equation (8) needs no learned critic.

Averaging over the current action, we recover the state-value V-trace operator $[\mathcal{V}_c^{\mathrm{s}} V]_t(x) := \mathbb{E}_{a\sim\mu}[[\mathcal{V}_c^{\mathrm{sa}} V]_t(x,a)]$ introduced by Espeholt et al. (2018) in the *untruncated* residual-ratio case ($\bar{\rho} = \infty$ in their notation). Now, writing $\mathcal{R}_c^\pi$ for the finite-horizon, next-action-indexed Retrace-form operator (Munos et al., 2016), the three operators satisfy pointwise

$$[\mathcal{B}(\mathcal{V}_c^{\mathrm{s}} V)]_t = [\mathcal{R}_c^\pi(\mathcal{B}V)]_t, \qquad [\mathcal{V}_c^{\mathrm{sa}} V]_t = \rho_t\big[(1-\lambda_t)[\mathcal{B}V]_t + \lambda_t[\mathcal{R}_c^\pi(\mathcal{B}V)]_t\big],$$

with $\lambda_t = c_t/\rho_t$ as above: a convex combination of a behavior and a Retrace backup, carried by the current-action change of measure ρ_t. Both are polynomial identities in the ratios and coefficients,

exact at any finite horizon rather than contraction or fixed-point results. Definitions, boundary cases, proofs and the dense-reward extension appear in Appendices B–J.3.

3 V-TRACE POLICY OPTIMIZATION

In Equation (2) the multiplier w_t is detached, so the active surrogate has coefficient $w_t\rho_t = \widetilde{\rho}_t$: the per-token score direction is the suffix trace, while the PPO mask tests ρ_t. Clipping the transformed trace $\hat{\rho}_t := \mathrm{sg}(\widetilde{\rho}_t/\rho_t)\rho_t$ instead — the variant we call V-TRACE PO-TC — gives the same unclipped score direction and differs only in which quantity the mask tests (Appendix M.8).

Equation (1) is the clearest mathematical form but should not be implemented as a literal division by a potentially tiny ρ_t. The backward scan of Equation (7), with resets at response boundaries, is linear in sequence length; under reuse the stored behavior denominator stays fixed while the learner ratio and trace are recomputed after every update.

The switching theorem is stated for terminal return R and applies to $\widehat{G}_c$ of Equation (8): exact ratios attached to the score, unclipped, and not yet divided by the random valid-token count Z. Proposition 2.2 centers each response's trace-score sum, so a baseline built from the other responses drops out: mean-only leave-one-out (Ahmadian et al., 2024) preserves the expected direction, and inclusive K-response mean subtraction as in GRPO (Shao et al., 2024) scales it by $(K-1)/K$ (Cor. C.3). Group standardization, the division by Z and PPO clipping are further transformations: the first breaks the corollary's identity, whereas subtracting the group mean alone only rescales it.

Appendix D gives finite-precision constructions, masks and safeguards for both variants, endpoint qualifications and complete pseudocode (Algorithm 1); Appendix J.3 extends the estimator to dense per-token rewards.

4 A SECOND MECHANISM UNDER SAMPLE REUSE

Sections 2–3 justify the suffix trace as a change of measure on the continuation. That argument fixes the behavior policy and uses each sample once. Sample reuse breaks its premise, and not only for suffix traces: once the learner has been updated on the batch, π_θ depends on the very action a_t it reweights, so $\mathbb{E}_\mu[\rho_t \mid \mathcal{F}_t] = 1$ — which needs π_θ fixed — no longer follows. This section shows that under sample reuse the same estimator acquires a second, distinct mechanism, and Section 6 shows that it is this second mechanism, not the change of measure, that tracks where the estimator helps.

Reuse makes the suffix deviations coherent. Suppose the learner is updated on a batch and then reweights that same batch. Response b contributed to that update along its own score direction $\sum_t s_t$ scaled by its advantage A_b, which suggests its token probabilities were pushed up when $A_b > 0$ and down when $A_b < 0$. The argument is heuristic—responses in a batch share an update, so a response's ratio also moves with the others' contributions—so we treat it as a hypothesis and measure it. If it holds, $\rho_s > 1$ along a positive-advantage response's suffix and $\rho_s < 1$ along a negative one's. Substituting into Equation (7), the suffix deviations share a sign and therefore *accumulate* rather than cancel. Writing $w_t = \widetilde{\rho}_t/\rho_t$ for the detached multiplier that V-TRACE PO applies outside its mask,

$$A_b > 0 \implies w_t > 1 \text{ (update amplified)}, \qquad A_b < 0 \implies w_t < 1 \text{ (update damped)}.$$

The predicted effect is large rather than marginal: at $\bar{c} = 1$ and $\epsilon = 1$, the 99th percentile (p99) of the suffix trace on the last of eight passes over a batch is 6.16, although the per-token log-ratio there averages only .0136: small deviations accumulate along the suffix (Appendix Table 5). When behavior and learner differ for reasons unrelated to the sampled advantage—a stale generator, or a hotter sampler—$\rho_s - 1$ is undirected with respect to A_b, the deviations cancel between positive- and negative-advantage responses, and the trace adds spread without separating the two. Only the difference between the two conditional means is informative. The prediction is specific: the asymmetry should track *whether* the mismatch was caused by the sample's own advantage, not *how large* the mismatch is.

Measurement. We log $\mathbb{E}[w \mid A > 0]$ and $\mathbb{E}[w \mid A < 0]$ as detached diagnostics in three of the regimes of Section 6: rollout reuse, a lagging generator and a hotter sampler. The prediction was

registered before the diagnostic existed, so these measurements test it rather than describe data already seen.

The result is a clean dissociation. Only when the batch is reused does w line up with the advantage, running above one on positive-advantage responses and below one on negative ones. A lagging generator and a hotter sampler (both *fresh sample* cases) produce essentially none of it, even though each puts μ and π *further* apart than sample reuse does. So the effect does not follow how far apart the two policies are, which is what the change of measure would predict; it follows whether the learner has already trained on the batch, and it strengthens with each further pass over it. Appendix M.11 reports the measurements.

5 Relation to prior work

Classical off-policy policy gradients distinguish action correction from state-distribution correction (Sutton et al., 1999; Degris et al., 2012). V-trace stabilizes off-policy value learning with truncated traces (Espeholt et al., 2018); Retrace provides a safe action-value return operator (Munos et al., 2016); and ACER combines multi-step critic targets, truncated actor ratios, bias correction, and a trust-region projection (Wang et al., 2017). These methods place a trace in a critic's value target and then apply a one-step actor ratio. Our state-action construction puts the trace directly on the per-token score, which on complete terminal-return responses removes the critic and makes its relation to both operators explicit.

Recent LLM objectives control different portions of the autoregressive ratio. GRPO and its variants retain token-local PPO structure (Shao et al., 2024; Ahmadian et al., 2024; Liu et al., 2025); GSPO aggregates ratios at response level (Zheng et al., 2025a); CTPO, MinPRO, and PNPO act on causal prefixes (Zhang et al., 2026b; Lei et al., 2026; Zhang et al., 2026a); and NFPO uses a fixed suffix window outside a separate local mask (Yoon et al., 2026). Stabilization and systems work addresses clipping asymmetry, batch filtering, stale actors, and asynchronous generation (Yu et al., 2025; Zheng et al., 2025b; Wu et al., 2025; Sheng et al., 2025; Fu et al., 2025). A separate line treats positive and negative examples asymmetrically by design (Arnal et al., 2025; Le Roux et al., 2025); the asymmetry of Section 4 is an emergent form of theirs, arriving through the trace rather than by design. V-Trace PO is otherwise complementary, leaving the prefix distribution at μ and changing only the continuation used to score the current action; Appendix E compares mechanisms in detail.

6 Experiments

Protocol. All experiments start from Qwen3-8B-Base (Yang et al., 2025) and train on MATH (Hendrycks et al., 2021). A fresh batch has 32 prompts and eight responses per prompt, sampled at temperature one unless stated otherwise. The stored sampler log-probability is the behavior denominator; reward-side and actor-loss KL terms are disabled. The primary endpoint, **Recent-150**, pools all 30 problems from each of five recent contests released through MathArena (Balunović et al., 2025; Dekoninck et al., 2026); canonical MATH-500 (Lightman et al., 2024) is reported separately and AIME 2024 is standalone and descriptive; no outcome-selected subset is used. The frozen model scores 5.00% on Recent-150, but the training signal is not sparse: response success averages 45.96% on the audited initial batches. Decoding, data provenance, hyperparameters and the statistical protocol are in Appendix M.

Every experiment reported here was preregistered: a factorial over reuse and clipping radius (Section 6.1); three ways of creating mismatch *without* reuse, together with a sweep of the truncation cap at one of them (Table 1); and an ablation of the mechanism (Section 6.2). All use the protocol above, with mean-only advantages: centered on the group mean and not divided by its standard deviation. The sequential-minibatch study, which ran earliest, instead divides by it, as GRPO does by default; it is marked wherever it appears, and no experiment mixes the two conventions, so the choice cannot confound any contrast reported here (Appendix M).

We fixed each experiment's design and analysis before running it. We did not fix the order: each experiment was chosen after seeing the last, so the series as a whole is exploratory. A paired sign-flip test on six seeds has only 2^6 possible sign assignments, so the smallest two-sided p-value it can return is $2/64 = .03125$; we write it as p throughout. Correcting for testing across experiments would

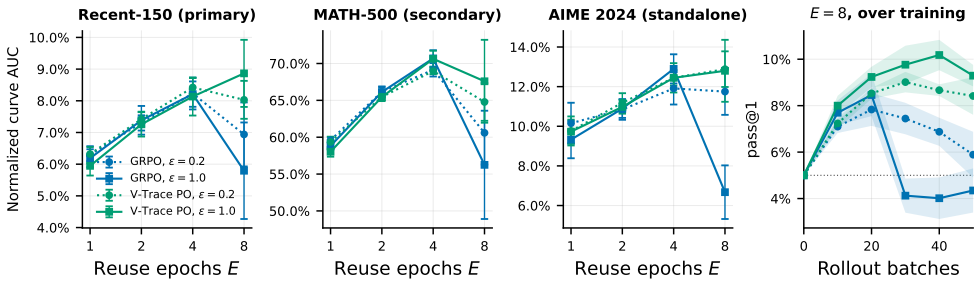

Figure 2: **Sweeping the number of passes E over each sampled batch.** First three panels: normalized curve AUC against E, at both trust-region widths and on all three endpoints, 95% Student-t intervals over six seeds. At $E = 8$, widening the trust region from $\epsilon = 0.2$ to $\epsilon = 1$ helps V-TRACE PO ($+0.83$) and hurts GRPO (-1.15); at lower reuse the radius changes little for either. Fourth panel follows $E = 8$ over training: both GRPO arms give their gains back after batch 20, the wide-radius one falling below the untrained model (dotted), while both V-TRACE PO arms climb for longer—to 10.2% at $\epsilon = 1$, the highest any arm reaches at that radius. Its shaded bands are ± 1 standard error of the mean over seeds, narrower than the 95% intervals in the other panels.

require a threshold below that, which no result could reach, so we report effect sizes with intervals where they bear on a claim and treat p-values as descriptive.

6.1 SUFFIX CORRECTION HELPS, AND MAINLY UNDER ROLLOUT REUSE

We ran a factorial crossing three arms, two clipping radii $\epsilon \in \{.2, 1\}$, reuse depth $E \in \{1, 2, 4, 8\}$ and six seeds (144 runs): GRPO at $\bar{c} = 0$, V-TRACE PO at $\bar{c} = 1$, and V-TRACE PO-TC (Appendix M.8). Each run consumes 50 fresh batches, with a sampled batch and its behavior log-probabilities held fixed through its E updates while the learner ratio and trace are recomputed. Normalized curve AUC is the trapezoidal mean of pass@1 over fresh batches 0–50, and training seed is the paired unit; the design is rollout-matched, not compute-matched.

The gap over GRPO grows monotonically with reuse at both radii (Figure 2), by $+0.328$ AUC points per doubling of E at $\epsilon = .2$ and $+0.999$ at $\epsilon = 1$ (both exact $p = .0625$), so widening the trust region amplifies the pattern. At $E = 8$, $\epsilon = 1$, GRPO reaches 5.793% and V-TRACE PO 8.865%, a gap of $+3.07$ points ($[0.67, 5.47]$, $p = .0625$); the MATH-500 gap is $+11.34$ ($[-1.2, 23.9]$, $p = .09375$) and the descriptive AIME 2024 gap $+6.13$ ($[3.6, 8.7]$, $p = .03125$). Two qualifications are visible in the figure. At the wide radius the estimator matches GRPO until reuse is aggressive (the $E = 1$ gap is -0.24), and the $E = 8$ gap opens largely because GRPO degrades on the last doubling, 8.211% at $E = 4$ to 5.793%, while V-TRACE PO moves 8.135% to 8.865%—its own best cell and the best of the 144 (Appendix Table 22).

The rest of this section asks what produces it.

6.2 WHICH MECHANISM IS RESPONSIBLE?

Section 4 gives two explanations that predict different things. The change of measure predicts a benefit whenever Q_t^μ and Q_t^π genuinely differ, however that difference arises; the asymmetry predicts one only when the mismatch was caused by the sample's own advantage, however large the mismatch is otherwise.

Mismatch without reuse. We ran three experiments at $E = 1$, where every response is consumed once: a frozen generator, a hotter sampler and sequential minibatches, each creating a mismatch between μ and π without any reuse. The frozen generator varies staleness over two orders of magnitude, moving mean $|\log \rho_t|$ more than tenfold, and the hotter sampler reaches three times the level seen under reuse. We measured no benefit in any of them; at the hottest sampler the gap is $+0.008$. We then swept the cap $\bar{c}$ through $\{0, .5, 1, 2\}$ at fixed mismatch, which raises the applied

p99 through 1.00, 1.10, 1.86 and 7.65 in turn; more correction did not help at any setting. Table 1 gives all four experiments.

These results are not all exactly zero: one of the sixteen cells gains $+0.181$ at all six seeds — the only cell positive in every pair, though not the largest — a sixteenth of the reuse effect, which we report and do not build on. In the two experiments where we logged the diagnostic there is no advantage-sign asymmetry at all: what the trace adds in these regimes is spread rather than a shift.

Table 1: **No benefit from suffix correction without sample reuse.** All at $E = 1$, so every response is used once. The first three experiments each create a mismatch between μ and π; the fourth holds the mismatch at the level of row two and varies the correction instead, so it is not an independent test. Each entry is that experiment's registered interaction over six paired seeds, with its exact two-sided sign-flip p: for rows one, two and four, the change in the V-TRACE PO-minus-GRPO gap across the row's mismatch levels, in Recent-150 curve AUC; for row three, the change across the clipping radius at fixed $M = 4$, in pass@1 at 400 updates, and the one study using group-standardized advantages. No row shows a reliable effect, and every plain contrast against GRPO lies between -0.68 and $+0.21$ curve-AUC points (row three, in pass@1: -0.92 and -0.19).

experiment	mean $\lvert \log \rho_t \rvert$	interaction	p
frozen generator, $K \leq 128$ updates	.0027 → .0333	$+0.019$	.96875
hotter sampler, $T_{\text{gen}} \leq 1.4$	→ .042	$+0.276$	.09375
sequential minibatches, $M = 4$	rises with position	$+0.72$	.34375
cap ladder at T_{gen}=1.2, $\bar{c} \in \{0, .5, 1, 2\}$	.027	$+0.319$	.3125

The numerical noise alone exhausts the variance budget. We also ran the cap sweep at $T_{\text{gen}} = 1.0$, and that control explains the rest. There behavior and target are the same weights, so the correction is zero in expectation and the only ratio deviation is numerical disagreement between the inference and training engines, mean $\lvert \log \rho_t \rvert = .009$ per token. The untruncated trace is a product over the remaining tokens, so it compounds that 1% disagreement into a p99 multiplier of 2.5 and costs 0.68 AUC points ($[-1.11, -0.25]$, $p = .0625$) where there is provably nothing to correct. The correction is already unprofitable before any real policy mismatch arrives, and no single-pass regime we construct shows a benefit. Shrinking the cap does not rescue the trade: whether any correction pays is decided at the margin, by a comparison that does not involve the horizon (Cor. J.3). Setting the horizon directly over 1–20 tokens finds no level that does, the best being $+0.06$ Recent-150 points over GRPO at exact $p = .75$ (Appendix M.9).

Ablating the asymmetry directly. Across the three regimes, the benefit follows the asymmetry rather than the size of the mismatch — the two run in almost opposite orders. That is still a correlation, so we manipulate the mechanism directly, inside the reuse setting at $E = 8$, $\epsilon = 1$, with six paired seeds. One arm, *shuffled*, rescales each response's multiplier so that its mean becomes another response's, destroying the coupling to the advantage while preserving token shape, length and gradient support; a second replaces the multiplier by constants keyed on $\text{sign}(A)$, with no trace. Because the rescaling adds noise as well as removing structure, its registered contrast is against GRPO rather than against the intact trace. The registered manipulation check failed: the rescaling removed 86% of the asymmetry rather than all of it, and the shuffled arm landed at neither of the two registered readings. The intact trace also applies some seven times the asymmetry of the other arms, so we read Table 2 as diagnostic rather than as a test. The reuse benefit replicates, and the two arms with matched applied asymmetry (0.408 and 0.414) sit within half a point of each other, above GRPO by $+1.39$ ($[0.21, 2.56]$) and $+0.94$ ($[-0.16, 2.03]$), with the intact trace highest. Descriptively, a constant pair carrying no trace recovers about half the benefit, so the trace is not the only route to the asymmetry. Constructions, registered readings and all five contrasts are in Appendix M.12.

Rollout budget, not update budget. These designs fix the number of fresh batches, so an $E = 8$ arm also takes eight times the actor updates. A matched-update comparison puts the $E = 8$ V-TRACE PO-TC recipe -1.89 Recent-150 and -7.61 MATH-500 points behind long-horizon $E = 1$ GRPO; only a twofold rollout reduction has support, on Recent-150 but not MATH-500, and only while spending four times the actor updates (Appendix L.6). We report against a fixed rollout budget because generation is the expensive side (Arnal et al., 2026).

Table 2: **Accuracy is consistent with the applied asymmetry rather than with its source.** Recent-150 curve AUC at $E = 8$, $\epsilon = 1$, six paired seeds.

arm	applied gap	AUC	mechanism
GRPO	0.000	5.51	none
fixed pair	0.408	6.90	constant, no trace
V-TRACE PO shuffled	0.414	6.45	trace, coupling destroyed
V-TRACE PO	2.857	8.19	trace, intact

7 LIMITATIONS

The change of measure that motivates the estimator is not what our experiments show makes it work: three designs create mismatch without reuse and none gives a benefit. We still report the derivation: it makes the suffix trace well posed, and finding no effect is consistent with a real correction smaller than its variance cost. Setting the horizon directly, from one to twenty tokens, finds no setting that pays either, though that result is flat rather than declining (Appendix M.9); and the ceiling we describe assumes the Q^π–Q^μ gap accrues evenly along the continuation, where a front-loaded gap would favor a short horizon.

Several things qualify the asymmetry account. The asymmetry appears under sample reuse and not with fresh samples, but that comparison cannot separate reuse from self-influence: a reused batch is by definition one the learner has already trained on. The shuffled arm, meant to destroy the link between the multiplier and the advantage, removed only 86% of it, so its registered gap to GRPO is biased upward by the asymmetry left in place and downward by the noise the rescaling adds. How many passes a batch has had is read off the order of log entries rather than recorded directly; on what that ordering calls the first pass over a fresh batch the asymmetry is already $+0.29$, where the account predicts none; and the diagnostic probes use one seed each. And while V-TRACE PO does produce an asymmetry, we do not show it is the right one: its size is whatever the trace yields, and we have not looked for other rules that would produce a calibrated asymmetry without hand-set constants.

Several assumptions limit the construction. Clipping, averaging the loss over valid tokens and the numerical safeguards all change the exact direction analyzed in Section 2, and suffix ratios cannot correct the prefix distribution, which stays at μ. Our only baseline is matched GRPO; the methods designed for this setting (Zheng et al., 2025a; Xi et al., 2025; Tian et al., 2026; Yoon et al., 2026) are compared mechanism by mechanism in Appendix E (Table 4) but not run. We study one model family and one task; broader settings would be needed to map where suffix correction is worth its variance.

8 CONCLUSION

V-TRACE PO attaches a coefficient-truncated suffix trace to the per-token score in a single reverse scan, and the truncation is not a free knob: each $\bar{c}$ defines exactly which continuation policy it targets. It helps when rollouts are reused rather than freshly sampled. On the last doubling of reuse a matched GRPO drops 29% below its $E = 4$ accuracy while V-TRACE PO gains 9% and reaches its own best, and at that depth widening ϵ from 0.2 to 1 helps V-TRACE PO where it hurts GRPO, so the correction is what buys robustness at a fixed rollout budget.

Two findings came out of analyzing V-TRACE PO empirically, and both reach past it. The first is a ceiling on suffix correction itself: under a uniform-accrual assumption, correcting further into the continuation adds variance in proportion to the tokens corrected while the bias it removes has diminishing returns, and no horizon between one and twenty tokens paid for itself at the one mismatch level we swept horizons over; separately, numerical disagreement between the inference and training engines is alone enough to make the correction cost accuracy. The second is that the benefit under reuse comes from a mechanism the derivation does not predict: reusing a batch makes the trace favor actions an earlier pass already favored, and accuracy tracks that asymmetry rather than the size of the mismatch. The trace is one way to produce it; controlling it directly is what we leave to future work.

AI USE STATEMENT

We used generative AI tools in four of the categories declared on the submission form: to aid and polish writing; for retrieval and discovery, chiefly literature-search triage; for research ideation and execution; and to draft sections of this paper. Concretely, AI assistance contributed to the design and preregistration of the experiments reported here, to the implementation of the training, evaluation and analysis code, to the orchestration and diagnosis of the runs, and to drafting the abstract and several sections, which the authors then revised. The mathematical development is the authors' own; AI was used to check it, including machine-verification of the trace identities and operator relations in double precision.

We did not treat generated prose, code, citations, or numerical summaries as authoritative. The authors inspected the cited sources and reviewed the manuscript and the experimental implementation. Every reported statistic is produced from a preregistered design by an analyzer that emits nothing if a declared data-completeness check fails and that records the outcome of every declared manipulation check; one of those checks failed, and Section 6.2 reports it as such. Every statistic is reproducible from frozen design and target manifests carrying a SHA-256 digest for each input; discrepancies found during review were corrected and the corrections recorded. The authors take responsibility for the final paper and results.

ETHICS STATEMENT

The experiments use public mathematical problem datasets and involve no human subjects or personal data. We report dataset provenance and selection criteria, avoid outcome-selected benchmark subsets, and keep Recent-150 separate from older benchmarks. The principal cost is accelerator use; we therefore reuse frozen checkpoints and responses whenever new generation is unnecessary.

REPRODUCIBILITY STATEMENT

The estimator, both clipping placements, and the matched and factorial experiments are specified in the main text. We release no separate code artifact, because none is needed to reproduce the method: Algorithm 1 gives the complete loss as framework-agnostic pseudocode, including the reverse scan, the response-boundary resets, the finite-precision safeguards and the tensor conventions, and it is a drop-in replacement for a conventional PPO/GRPO loss in any reinforcement-learning codebase. Appendix D states the conventions it assumes. For the theoretical claims, Appendices F and G give complete proofs, and the coverage and support conditions they require, together with the uniform-accrual assumption of Corollary J.3, are stated where they are used rather than left implicit. For the experiments, Appendix M gives the model and dataset identifiers, the evaluation-suite construction and its exclusion criteria, every hyperparameter, the advantage convention used by each study, the statistical protocol, and seed-level results; Appendix M.5 gives the run and evaluation inventories, the fail-closed validation criteria, and the append-only provenance records for the infrastructure repairs, so that the audit can be checked rather than taken on trust.

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.

Charles Arnal, Gaetan Narozniak, Vivien Cabannes, Yunhao Tang, Julia Kempe, and Rémi Munos. Asymmetric REINFORCE for off-policy reinforcement learning: Balancing positive and negative rewards. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL `https://arxiv.org/abs/2506.20520`. arXiv:2506.20520.

Charles Arnal, Vivien Cabannes, Taco Cohen, Julia Kempe, and Rémi Munos. Efficient RL training for LLMs with experience replay. arXiv preprint arXiv:2604.08706, 2026.

Mislav Balunović, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin Vechev. MathArena: Evaluating LLMs on uncontaminated math competitions. In *Proceedings of the Neural Information*

Processing Systems Track on Datasets and Benchmarks, 2025. URL https://arxiv.org/abs/2505.23281. arXiv:2505.23281.

Taco Cohen, David W. Zhang, Kunhao Zheng, Yunhao Tang, Rémi Munos, and Gabriel Synnaeve. Soft policy optimization: Online off-policy RL for sequence models. arXiv preprint arXiv:2503.05453, 2025.

Thomas Degris, Martha White, and Richard S. Sutton. Off-policy actor-critic. In *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland, UK, 2012. URL https://arxiv.org/abs/1205.4839.

Jasper Dekoninck, Nikola Jovanović, Tim Gehrunger, Kári Rögnvaldsson, Ivo Petrov, Chenhao Sun, and Martin Vechev. Beyond benchmarks: MathArena as an evaluation platform for mathematics with LLMs, 2026. URL https://arxiv.org/abs/2605.00674.

Lasse Espeholt, Hubert Soyer, Rémi Munos, Karen Simonyan, Vlad Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, et al. IMPALA: Scalable distributed deep-RL with importance weighted actor-learner architectures. In *International Conference on Machine Learning*, pp. 1407–1416. PMLR, 2018.

Wei Fu, Jiaxuan Gao, Xujie Shen, Chen Zhu, Zhiyu Mei, Chuyi He, Shusheng Xu, Guo Wei, Jun Mei, Jiashu Wang, Tongkai Yang, Binhang Yuan, and Yi Wu. AReaL: A large-scale asynchronous reinforcement learning system for language reasoning. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://arxiv.org/abs/2505.24298v5. arXiv:2505.24298v5.

Anna Harutyunyan, Marc G. Bellemare, Tom Stepleton, and Rémi Munos. Q(λ) with off-policy corrections. In *International Conference on Algorithmic Learning Theory*, pp. 305–320. Springer, 2016.

Horace He and Thinking Machines Lab. Defeating nondeterminism in LLM inference. Thinking Machines Lab: Connectionism, 2025. URL https://thinkingmachines.ai/blog/defeating-nondeterminism-in-llm-inference/.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021. URL https://arxiv.org/abs/2103.03874. arXiv:2103.03874.

Jiawei Huang and Nan Jiang. From importance sampling to doubly robust policy gradient. In *International Conference on Machine Learning*, pp. 4434–4443. PMLR, 2020.

Luke J. Huang, Zhuoyang Zhang, Qinghao Hu, Shang Yang, and Song Han. Stable asynchrony: Variance-controlled off-policy RL for LLMs. arXiv preprint arXiv:2602.17616v2, 2026. URL https://arxiv.org/abs/2602.17616v2.

Ehsan Imani, Eric Graves, and Martha White. An off-policy policy gradient theorem using emphatic weightings. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://arxiv.org/abs/1811.09013.

Maxwell Jia. AIME 2024 dataset. Hugging Face dataset, 2024. URL https://huggingface.co/datasets/Maxwell-Jia/AIME_2024.

Tadashi Kozuno, Yunhao Tang, Mark Rowland, Rémi Munos, Steven Kapturowski, Will Dabney, Michal Valko, and David Abel. Revisiting peng's Q(λ) for modern reinforcement learning. In *International Conference on Machine Learning*, pp. 5794–5804. PMLR, 2021.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626, 2023. doi: 10.1145/3600006.3613165.

Nicolas Le Roux, Marc G. Bellemare, Jonathan Lebensold, Arnaud Bergeron, Joshua Greaves, Alex Fréchette, Carolyne Pelletier, Eric Thibodeau-Laufer, Sándor Toth, and Sam Work. Tapered off-policy REINFORCE: Stable and efficient reinforcement learning for LLMs. arXiv preprint arXiv:2503.14286, 2025.

Shiye Lei, Zhihao Cheng, and Dacheng Tao. A step back: Prefix importance ratio stabilizes policy optimization. arXiv preprint arXiv:2601.22718v1, 2026. URL https://arxiv.org/abs/2601.22718v1.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2305.20050. arXiv:2305.20050.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding R1-Zero-like training: A critical perspective. In *Proceedings of the 2nd Conference on Language Modeling*, 2025. URL https://arxiv.org/abs/2503.20783v2. arXiv:2503.20783v2.

Alberto Maria Metelli, Matteo Papini, Nico Montali, and Marcello Restelli. Importance sampling techniques for policy optimization. *Journal of Machine Learning Research*, 21(141):1–75, 2020.

MiniMax. MiniMax-M1: Scaling test-time compute efficiently with lightning attention. arXiv preprint arXiv:2506.13585v1, 2025. URL https://arxiv.org/abs/2506.13585v1.

Rémi Munos, Tom Stepleton, Anna Harutyunyan, and Marc G. Bellemare. Safe and efficient off-policy reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Doina Precup, Richard S. Sutton, and Satinder P. Singh. Eligibility traces for off-policy policy evaluation. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pp. 759–766. Morgan Kaufmann, 2000.

Penghui Qi, Xiangxin Zhou, Zichen Liu, Tianyu Pang, Chao Du, Min Lin, and Wee Sun Lee. Rethinking the trust region in LLM reinforcement learning. arXiv preprint arXiv:2602.04879v3, 2026. URL https://arxiv.org/abs/2602.04879v3.

Mark Rowland, Will Dabney, and Rémi Munos. Adaptive trade-offs in off-policy learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 34–44. PMLR, 2020.

Simon Schmitt, Matteo Hessel, and Karen Simonyan. Off-policy actor-critic with shared experience replay. In *International Conference on Machine Learning*, pp. 8545–8554. PMLR, 2020.

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

Sheikh Shafayat, Fahim Tajwar, Ruslan Salakhutdinov, Jeff Schneider, and Andrea Zanette. Can large reasoning models self-train?, 2025. URL https://arxiv.org/abs/2505.21444.

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

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. HybridFlow: A flexible and efficient RLHF framework. In *Proceedings of the Twentieth European Conference on Computer Systems (EuroSys)*, 2025. URL https://arxiv.org/abs/2409.19256. arXiv:2409.19256.

Richard S. Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, volume 12, 1999.

Yunhao Tang, Michal Valko, and Rémi Munos. Taylor expansion policy optimization. In *International Conference on Machine Learning*, pp. 9397–9406. PMLR, 2020.

Yunhao Tang, Tadashi Kozuno, Mark Rowland, Anna Harutyunyan, Rémi Munos, Bernardo Avila Pires, and Michal Valko. DoMo-AC: Doubly multi-step off-policy actor-critic algorithm. In *International Conference on Machine Learning*, pp. 33657–33673. PMLR, 2023.

Yunhao Tang, Taco Cohen, David W. Zhang, Michal Valko, and Rémi Munos. RL-finetuning LLMs from on- and off-policy data with a single algorithm. arXiv preprint arXiv:2503.19612, 2025.

Minghao Tian, Yunfei Xie, and Chen Wei. How off-policy can GRPO be? Mu-GRPO for efficient LLM reinforcement learning. arXiv preprint arXiv:2605.17570v1, 2026. URL https://arxiv.org/abs/2605.17570v1.

Ziyu Wang, Victor Bapst, Nicolas Heess, Volodymyr Mnih, Rémi Munos, Koray Kavukcuoglu, and Nando de Freitas. Sample efficient actor-critic with experience replay. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1611.01224.

Bo Wu, Sid Wang, Yunhao Tang, Jia Ding, Eryk Helenowski, Liang Tan, Tengyu Xu, Tushar Gowda, Zhengxing Chen, Chen Zhu, et al. LlamaRL: A distributed asynchronous reinforcement learning framework for efficient large-scale LLM training. arXiv preprint arXiv:2505.24034, 2025.

Zhiheng Xi, Xin Guo, Yang Nan, Enyu Zhou, Junrui Shen, Wenxiang Chen, Jiaqi Liu, Jixuan Huang, Zhihao Zhang, Honglin Guo, Xun Deng, Zhikai Lei, Miao Zheng, Guoteng Wang, Shuo Zhang, Peng Sun, Rui Zheng, Hang Yan, Tao Gui, Qi Zhang, and Xuanjing Huang. BAPO: Stabilizing off-policy reinforcement learning for LLMs via balanced policy optimization with adaptive clipping. arXiv preprint arXiv:2510.18927v1, 2025. URL https://arxiv.org/abs/2510.18927v1.

Jianhao Yan, Yafu Li, Zican Hu, Zhi Wang, Ganqu Cui, Xiaoye Qu, Yu Cheng, and Yue Zhang. Learning to reason under off-policy guidance. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://arxiv.org/abs/2504.14945v5. arXiv:2504.14945v5.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. arXiv preprint arXiv:2505.09388, 2025. URL https://arxiv.org/abs/2505.09388.

Feng Yao, Liyuan Liu, Dinghuai Zhang, Chengyu Dong, Jingbo Shang, and Jianfeng Gao. Your efficient RL framework secretly brings you off-policy RL training. https://fengyao.notion.site/off-policy-rl, August 2025. URL https://fengyao.notion.site/off-policy-rl. Web article.

Deokgyu Yoon, Hyungkyu Kang, Joongkyu Lee, Byeongchan Kim, Gyungin Shin, Sungrae Park, and Min-hwan Oh. Multi-step likelihood-ratio correction for reinforcement learning with verifiable rewards. arXiv preprint arXiv:2605.20865v1, 2026. URL https://arxiv.org/abs/2605.20865v1.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. DAPO: An open-source LLM reinforcement learning system at scale. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://arxiv.org/abs/2503.14476. arXiv:2503.14476.

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. arXiv preprint arXiv:2608.01418v1, 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:2605.07331v1, 2026b. URL https://arxiv.org/abs/2605.07331v1.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. arXiv preprint arXiv:2507.18071, 2025a.

Haizhong Zheng, Jiawei Zhao, and Beidi Chen. Prosperity before collapse: How far can off-policy RL reach with stale data on LLMs? arXiv preprint arXiv:2510.01161v2, 2025b. URL https://arxiv.org/abs/2510.01161v2.

A THE ANATOMY OF OFF-POLICY AUTOREGRESSIVE GRADIENTS

Notation guide. Formal definitions and boundary conventions follow in this appendix.

Symbol	Meaning
ρ_t, $\rho_{i:j}$	Target-to-behavior action ratio and its product from indices i through j.
c_t, λ_t, $c_{i:j}$	Trace coefficient, normalized continuation coefficient $\lambda_t = c_t/\rho_t$ (with the stated zero convention), and the coefficient product.
$\widetilde{\rho}_t$, $\hat{\rho}_t$, w_t	Cumulative suffix trace, stop-gradient transformed ratio $\hat{\rho}_t = \mathrm{sg}(\widetilde{\rho}_t/\rho_t)\rho_t$, and detached continuation multiplier $w_t = \mathrm{sg}(\widetilde{\rho}_t/\rho_t)$; when $\rho_t = 0$ both use the equivalent division-free form, giving $w_t = 1$ and $\hat{\rho}_t = 0$.
$\mathcal{V}_c^{\mathrm{sa}}$, $\mathcal{V}_c^{\mathrm{s}}$	State-action V-trace map $\mathbb{V} \to \mathbb{Q}$ and its behavior-action-marginal state-value map $\mathbb{V} \to \mathbb{V}$.
$\mathcal{R}_c^{\pi}$, $\mathcal{B}$, $\mathcal{T}^{\pi}$	Finite-horizon next-action Retrace-form operator, terminal-aware state-to-action lift, and target-policy Bellman backup.

A.1 FINITE-HORIZON SETUP

The initial state is sampled from a common policy-independent law, actions are sampled at $t = 0, \dots, T-1$, and the trajectory ends with an integrable terminal reward R. The state x_t contains the prompt and the complete observable history before action a_t; in language generation, $x_{t+1} = (x_t, a_t)$. We use a fixed behavior policy μ and a target policy $\pi = \pi_\theta$. Their state distributions are d_t^{μ} and d_t^{π}. For any policy ν, define restart values

$$V_t^{\nu}(x) := \mathbb{E}_{\nu}[R \mid x_t = x], \qquad Q_t^{\nu}(x, a) := \mathbb{E}_{\nu}[R \mid x_t = x, a_t = a],$$

where ν in Q_t^{ν} governs only actions strictly after the fixed action a_t. In particular, once the final action is fixed, $Q_{T-1}^{\nu}(x, a)$ is independent of ν.

For exposition, the state and action spaces are finite, the environment and reward law are policy independent, and π_θ is continuously differentiable. We assume the target is covered by the behavior policy at every behavior-reachable history:

$$\pi(a \mid x) > 0 \quad \Longrightarrow \quad \mu(a \mid x) > 0.$$

Let $\mathcal{X}_t^{\mu} := \{x : d_t^{\mu}(x) > 0\}$, $\mathcal{A}_{\mu}(x) := \{a : \mu(a \mid x) > 0\}$, and $\mathcal{Z}_t^{\mu} := \{(x, a) : x \in \mathcal{X}_t^{\mu},\ a \in \mathcal{A}_{\mu}(x)\}$. Ratios and conditional identities are understood on these behavior-supported sets. Identities stated conditionally on a behavior-supported pair $(x, a) \in \mathcal{Z}_t^{\mu}$, such as the switching identity of Theorem C.1, remain valid when $\pi(a \mid x) = 0$; the current ratio is then $\rho_t(x, a) = 0$. For score-function identities, we additionally work at an interior parameter and assume that $\pi_\theta(a \mid x) > 0$ on the behavior support throughout a neighborhood of that parameter (Sutton et al., 1999); the finite trajectory sum can then be differentiated term by term. The finite-space assumption is expository: at fixed finite horizon the identities extend to measurable spaces whenever the corresponding Radon–Nikodym derivatives and expectations exist. We make no infinite-horizon claim. Define

$$\rho_t := \frac{\pi(a_t \mid x_t)}{\mu(a_t \mid x_t)}, \qquad \rho_{i:j} := \prod_{k=i}^{j} \rho_k, \qquad s_t := \nabla_\theta \log \pi_\theta(a_t \mid x_t),$$

with empty products equal to one. The behavior policy and its induced state distributions remain fixed during differentiation.

A.2 PREFIX, CURRENT, AND SUFFIX RATIOS

To see what can be corrected without paying for a full trajectory product, consider a single score term at time t. Factor the trajectory into the prefix before t, the current action, and the suffix after t. Applying importance ratios to different factors gives the four exact identities in Proposition 2.1. Prefix ratios $\rho_{0:t-1}$ change the distribution of prefix states at which the score is evaluated, $d_t^{\mu} \to d_t^{\pi}$; ρ_t changes the sampled action law $\mu(\cdot \mid x_t) \to \pi(\cdot \mid x_t)$; and suffix ratios $\rho_{t+1:T-1}$ change the continuation target $Q_t^{\mu} \to Q_t^{\pi}$. Appendix F gives the proof.

Equation (3), summed over t, yields the exact full-trajectory policy-gradient identity

$$\nabla_\theta \mathbb{E}_{\pi_\theta}[R] = \mathbb{E}_\mu \left[R\rho_{0:T-1} \sum_{t=0}^{T-1} s_t \right]. \tag{10}$$

Equation (10) is an identity between expectations; the sample quantity inside it is the corresponding unbiased estimator of the on-policy gradient from behavior trajectories. By contrast, the current-token direction is exactly the gradient of the fixed-behavior surrogate

$$\mathcal{J}_{\mathrm{tok}}(\theta;\mu) := \sum_{t=0}^{T-1} \mathbb{E}_{x_t \sim d_t^\mu} \left[\sum_{a \in \mathcal{A}_\mu(x_t)} \pi_\theta(a \mid x_t) Q_t^\mu(x_t, a) \right]. \tag{11}$$

namely

$$\nabla_\theta \mathcal{J}_{\mathrm{tok}}(\theta;\mu) = \mathbb{E}_\mu \left[R \sum_{t=0}^{T-1} \rho_t s_t \right]. \tag{12}$$

Being an exact gradient of this surrogate does not imply that the surrogate equals the on-policy return, shares its optima under a shared neural parameterization, or guarantees improvement after a finite update.

The decomposition isolates a useful design axis. The current ratio is what changes the sampled action from μ to π; dropping it loses even the fixed-behavior surrogate interpretation. Correcting the entire prefix pays for products over the past. We instead ask how much of the *future* continuation should be corrected, while deliberately retaining $x_t \sim d_t^\mu$.

B SUFFIX-TRACE POLICY OPTIMIZATION

B.1 CUMULATIVE SUFFIX TRACE

The previous section leaves one degree of freedom: how far beyond the current token should the change of measure extend? V-TRACE PO answers with a trace whose coefficients determine an adaptive correction horizon.

For the target analysis, let each trace coefficient be a state-action function $c_t = c_t(x_t, a_t)$, held fixed in θ (that is, evaluated and detached at the current iterate), satisfying

$$0 \le c_t \le \rho_t. \tag{13}$$

When $\rho_t > 0$, write $\lambda_t = c_t/\rho_t$; when $\rho_t = 0$, set $\lambda_t = 0$, which is consistent with $c_t = 0$. Hence $c_t = \rho_t \lambda_t$ and $\lambda_t \in [0, 1]$. We use $c_{i:j} := \prod_{k=i}^{j} c_k$, again with empty products equal to one.

Definition B.1 (Cumulative suffix trace and sample direction) *Define*

$$\widetilde{\rho}_t := \rho_t + \sum_{s=t+1}^{T-1} c_{t:s-1}(\rho_s - 1). \tag{14}$$

The unclipped suffix-trace sample direction is

$$\widehat{G}_c := R \sum_{t=0}^{T-1} \widetilde{\rho}_t s_t. \tag{15}$$

The trace is computed in one reverse pass:

$$\widetilde{\rho}_{T-1} = \rho_{T-1}, \qquad \widetilde{\rho}_t = \rho_t + c_t(\widetilde{\rho}_{t+1} - 1), \quad t < T-1. \tag{16}$$

Equivalently, $\widetilde{\rho}_t = \rho_t[(1-\lambda_t) + \lambda_t \widetilde{\rho}_{t+1}]$, which also proves nonnegativity by backward induction. The recursion is a pathwise algebraic rearrangement of Equation (14) and needs no probabilistic assumption; adaptedness of the coefficients is used only for the conditional-expectation results below. Two useful trace families are

$$c_t = \min\{\bar{c}, \rho_t\} \quad \text{and} \quad c_t = \lambda \rho_t,$$

for a threshold $\bar{c} \ge 0$ or constant $\lambda \in [0, 1]$. The threshold family adapts to each observed ratio. The constant-λ family instead has a random-horizon interpretation: it averages partial suffix products with geometrically decaying weights, plus a final atom for never switching before termination.

Proposition B.2 (Exact endpoints) *If $c_t = 0$, then $\widetilde{\rho}_t = \rho_t$; every term of Equation* (14) *beyond the first carries c_t as a factor. If $c_j = \rho_j$ on the same suffix, then $\widetilde{\rho}_t = \rho_{t:T-1}$.*

Proposition B.3 (Tail moderation under a bounded cap) *For any adapted coefficients with $0 \le c_j \le \bar{c}$ — in particular the threshold family — and any $\bar{c} \ge 0$,*

$$|\widetilde{\rho}_t - \rho_t| \le \sum_{s=t+1}^{T-1} \bar{c}^{s-t} |\rho_s - 1|.$$

Proof. The condition $0 \le c_j \le \bar{c}$ gives $0 \le c_{t:s-1} \le \bar{c}^{s-t}$, since $c_{t:s-1}$ has $s-t$ factors. The triangle inequality applied to $\widetilde{\rho}_t - \rho_t = \sum_{s>t} c_{t:s-1}(\rho_s - 1)$ in Equation (14) then gives the bound. □

The signed deviation $\widetilde{\rho}_t - \rho_t$ can take either sign, because each term $\rho_s - 1$ enters with its own sign: the trace both amplifies and damps, and Section 4 relies on downward deviations as much as upward ones. Ratio deviations at distant s enter at most additively when $\bar{c} = 1$ and decay geometrically when $\bar{c} < 1$, instead of being multiplied into a full suffix product. The bound does not bound $\widetilde{\rho}_t$ itself: the local ratios and the finite sum can still be arbitrarily large.

At the second endpoint, the correction is over the complete suffix, not the complete trajectory: the prefix x_t remains distributed according to d_t^μ. For $c_j = \lambda \rho_j$, the exact finite-horizon expansion is

$$\widetilde{\rho}_t = (1-\lambda) \sum_{k=0}^{T-t-2} \lambda^k \rho_{t:t+k} + \lambda^{T-t-1} \rho_{t:T-1}. \tag{17}$$

The last term is the finite-horizon atom corresponding to never switching before termination; omitting it would leave a sub-probability mixture. A visual expansion appears in Appendix K.

B.2 State-action V-trace and the critic-free realization

We now define the operator used by the estimator. Let $\mathbb{V}$ be the time-indexed state functions $V = (V_0, \dots, V_T)$ on $\mathcal{X}_t^\mu$, and let $\mathbb{Q}$ be the corresponding state-action functions on $\mathcal{Z}_t^\mu$. Define the terminal-aware state-to-action lift

$$[\mathcal{B}V]_t(x,a) := \begin{cases} \mathbb{E}[V_{t+1}(x_{t+1}) \mid x_t = x, a_t = a], & t < T-1, \\ Q_{T-1}^\mu(x,a), & t = T-1, \end{cases}$$

The terminal value is policy-independent because no future action remains. Thus $\mathcal{B} : \mathbb{V} \to \mathbb{Q}$. The terminal coordinate V_T is left unconstrained and is never used: both $\mathcal{B}$ and the residual δ_s^V below special-case $s = T-1$ and consume the sampled return directly. The main text folds this into the single convention $V_T := R$, which makes its one uniform residual display valid at $s = T-1$ as well.

Define the terminal-aware residual

$$\delta_s^V := \begin{cases} \rho_s V_{s+1}(x_{s+1}) - V_s(x_s), & s < T-1, \\ \rho_{T-1} R - V_{T-1}(x_{T-1}), & s = T-1. \end{cases}$$

Its baseline term is $-V_s(x_s)$ rather than $-\rho_s V_s(x_s)$; this is the choice that yields both the local-ratio endpoint and the terminal-return realization of Proposition B.6.

Definition B.4 (Finite-horizon state-action V-trace) *The state-action V-trace operator $\mathcal{V}_c^{\mathrm{sa}} : \mathbb{V} \to \mathbb{Q}$ maps V to the unnormalized, action-conditioned target*

$$[\mathcal{V}_c^{\mathrm{sa}} V]_t(x,a) = V_t(x) + \mathbb{E}_\mu \left[\sum_{s=t}^{T-1} c_{t:s-1} \delta_s^V \,\middle|\, x_t = x, a_t = a \right], \qquad 0 \le t < T, \tag{18}$$

whose terminal case is

$$[\mathcal{V}_c^{\mathrm{sa}} V]_{T-1}(x,a) = \rho_{T-1}(x,a) Q_{T-1}^\mu(x,a). \tag{19}$$

The output stays conditioned on the current action and retains its importance ratio. The remaining time steps admit the following efficient evaluation rule.

Proposition B.5 (Backward recursion) *For $V \in \mathbb{V}$ and $t < T-1$, writing $\rho_t = \rho_t(x_t, a_t)$ and $c_t = c_t(x_t, a_t)$,*

$$[\mathcal{V}_c^{\mathrm{sa}} V]_t(x_t, a_t) = \rho_t[\mathcal{B}V]_t(x_t, a_t) + c_t \mathbb{E}_\mu[[\mathcal{V}_c^{\mathrm{sa}} V]_{t+1}(x_{t+1}, a_{t+1}) - V_{t+1}(x_{t+1}) \mid x_t, a_t]\,. \quad (20)$$

The inner expectation averages over the transition to x_{t+1} and over the next behavior action $a_{t+1} \sim \mu(\cdot \mid x_{t+1})$. That behavior-action marginalization is the one place the tower property is used: the residual tail $\sum_{s>t} c_{t+1:s-1}\delta_s^V$ is measurable with respect to the history after a_{t+1}, so conditioning it on (x_{t+1}, a_{t+1}) turns it into the next state–action target $[\mathcal{V}_c^{\mathrm{sa}} V]_{t+1} - V_{t+1}$.

The placement of the current ratio and baseline in Equation (18) is deliberate. At $c = 0$ the recursion gives the local target $\rho_t[\mathcal{B}V]_t$; and for any admissible c, when applied to the behavior value V^μ, the sum of defining residuals telescopes pathwise into $R\widetilde{\rho}_t$, the observed terminal return weighted by the cumulative suffix trace (Proposition B.6). Other action-conditioned lifts are possible, but need not share both properties. Appendix G proves the recursion and gives a second, telescoped non-recursive form.

Proposition B.6 (Terminal-return realization) *For a complete behavior trajectory — substituting the sampled return R for each $V_s^\mu(x_s)$ in Equation* (18)*, which the tower property permits because each multiplying factor is measurable at its own history (Appendix F) —*

$$R\left[c_{t:T-2}\rho_{T-1} + \sum_{s=t}^{T-2} c_{t:s-1}(\rho_s - c_s)\right] = R\widetilde{\rho}_t$$

pathwise, and

$$\mathbb{E}_\mu[R\widetilde{\rho}_t \mid x_t, a_t] = [\mathcal{V}_c^{\mathrm{sa}} V^\mu]_t(x_t, a_t).$$

Critic-free realization. On a complete terminal-reward trajectory, the residual-sum Monte Carlo realization of state-action V-trace at V^μ telescopes pathwise to $R\widetilde{\rho}_t$. The sampled return therefore supplies the multi-step target without a learned critic. A live truncated trajectory still requires a boundary estimate.

Thus no parametric value or Q function is learned, stored, or bootstrapped for complete terminal-reward trajectories. The observed return is a Monte Carlo sample of each behavior continuation value, and the action-conditioned residual sum then telescopes pathwise into $R\widetilde{\rho}_t$. This exchanges critic approximation and computation for Monte Carlo variance. If an unroll stops before the actual terminal boundary, an appropriate boundary estimator—often a learned critic—is required rather than simply dropping the remaining continuation; an inaccurate estimator can itself add bootstrap bias.

The endpoints are algebraically clear. The remaining question is what continuation policy an intermediate trace represents.

C WHICH CONTINUATION DOES THE TRACE TARGET?

For the two endpoints, the answer is Q_t^μ and Q_t^π. Intermediate traces admit an equally concrete interpretation, provided that the switching time is defined carefully. Conditional on a current pair (x_t, a_t), the current action has already been changed from μ to π by ρ_t. After that action, toss a coin with success probability $\lambda_t(x_t, a_t)$. Conditional on the visited state-action pairs, these auxiliary coins are independent of one another and of future environment randomness. On failure, draw all later actions from μ. On success, draw a_{t+1} from π, then repeat the decision after that next action. There is no switch decision after a_{T-1}.

Let $Q_t^c(x, a)$ be the action value of this complete switching process started at $(x_t, a_t) = (x, a)$ with no earlier switch decision having failed; its superscript indexes the trace rule rather than a policy. After the first failure every later action is drawn from μ, so the remaining value is Q^μ; this is why the failure branch of the recursion below is Q_t^μ and not another Q^c. Equivalently, set the terminal value

and, for $t < T-1$, recurse as

$$Q^c_{T-1}(x,a) := Q^\mu_{T-1}(x,a),$$
$$Q^c_t(x,a) := (1-\lambda_t(x,a))Q^\mu_t(x,a) + \lambda_t(x,a)\mathbb{E}\left[\sum_{a'\in\mathcal{A}_\mu(x_{t+1})} \pi(a' \mid x_{t+1})Q^c_{t+1}(x_{t+1},a') \,\middle|\, x_t = x, a_t = a\right], \quad (21)$$

The expectation is the branch that forces the next action to follow π; the outer mixture with Q^μ_t is applied exactly once. This placement prevents an off-by-one switch. Figure 1 in the main text depicts the post-action switch timing.

Theorem C.1 (Finite-horizon switching target) *Assume Equation* (13). *Then for every* $t \in \{0,\dots,T-1\}$*, every behavior-supported state–action pair* (x,a) *obeys*

$$\boxed{\mathbb{E}_\mu[R\widetilde{\rho}_t \mid x_t = x, a_t = a] = \rho_t(x,a)Q^c_t(x,a).}$$

When $\rho_t(x,a) = 0$*, Equation* (13) *forces* $c_t(x,a) = 0$*, so* $\widetilde{\rho}_t = \rho_t + c_t(\widetilde{\rho}_{t+1} - 1) = 0$ *pathwise and the left-hand side vanishes; the right-hand side vanishes through its explicit factor* $\rho_t(x,a)$*, and only there is the convention* $\lambda_t := 0$ *used, to keep* $Q^c_t(x,a)$ *well defined. Consequently, using the coverage assumption of Section A to write the sum over* $\mathcal{A}_\mu(x_t)$ *as an expectation under* π*, and writing* $G_c(\pi;\mu) := \mathbb{E}_\mu[\widehat{G}_c]$*,*

$$G_c(\pi;\mu) = \sum_{t=0}^{T-1} \mathbb{E}_{\substack{x_t\sim d^\mu_t \\ a_t\sim\pi(\cdot|x_t)}} [Q^c_t(x_t,a_t)s_t]. \quad (22)$$

Proposition C.2 (Conditional trace normalization and score centering) *Let* $\mathcal{F}_t$ *denote the prompt and complete history immediately before* a_t *is sampled, so* x_t *is* $\mathcal{F}_t$*-measurable and* $\mathcal{F}_{t+1}$ *also contains* a_t*. Under the exact-ratio and support assumptions of Section A, for every decision time,*

$$\mathbb{E}_\mu[\widetilde{\rho}_t \mid \mathcal{F}_t] = 1, \quad (23)$$
$$\mathbb{E}_\mu[\widetilde{\rho}_t s_t \mid \mathcal{F}_t] = 0. \quad (24)$$

The result holds for any integrable adapted trace rule $c_t = c_t(x_t,a_t)$*; the stronger condition* $0 \le c_t \le \rho_t$ *is needed for the switching interpretation, not for these two centering identities.*

Corollary C.3 (Mean-only RLOO and inclusive group centering) *Conditional on a prompt* $\mathcal{P}$*, let* $K \ge 2$ *behavior trajectories be independent, with terminal rewards* R_i*, and define*

$$H_i := \sum_{t=0}^{T-1} \widetilde{\rho}_{i,t}s_{i,t}, \qquad B_{-i} := \frac{1}{K-1}\sum_{j\ne i} R_j, \qquad \bar{R} := \frac{1}{K}\sum_{j=1}^{K} R_j.$$

Then

$$\mathbb{E}_\mu[(R_i - B_{-i})H_i \mid \mathcal{P}] = \mathbb{E}_\mu[R_iH_i \mid \mathcal{P}], \quad (25)$$
$$\mathbb{E}_\mu[(R_i - \bar{R})H_i \mid \mathcal{P}] = \frac{K-1}{K}\mathbb{E}_\mu[R_iH_i \mid \mathcal{P}]. \quad (26)$$

Thus a mean-only RLOO baseline, with both the baseline and the computed trace coefficient held fixed in the score attachment, preserves the expected unclipped trace-score direction. Inclusive mean subtraction has the same direction but scales it by $(K-1)/K$*; multiplying by* $K/(K-1)$ *restores the RLOO scale. Both statements follow from Equation* (24)*, which by the tower property makes* $\mathbb{E}_\mu[H_i \mid \mathcal{P}] = 0$*, and from the independence of the* K *trajectories given the prompt: a baseline built from the other responses is then uncorrelated with* H_i *and drops out, while the inclusive mean retains only its own* $1/K$ *share.*

The backward-induction proof in Appendix H.3 uses the conditional centering identity $\mathbb{E}_\mu[\widetilde{\rho}_{t+1} - 1 \mid \mathcal{F}_{t+1}] = 0$. This result is deliberately narrower than practical GRPO. In particular, division by a realized group standard deviation that includes R_i introduces a random factor statistically coupled

to trajectory i and is not covered by Equations (25)– (26). PPO clipping and data-dependent token normalization are additional transformations. Therefore, when group centering is intended purely as a control variate without changing the expected unclipped trace-score direction, the theoretically clean choice is mean-only RLOO (or rescaled inclusive mean centering), without group standard deviation normalization. Corollary C.3 is stated with a single horizon T shared by the K trajectories; responses of different lengths are covered by reading T as the padded horizon. For a bounded stopping time, the same proof applies on a padded space when token validity is known before its action is sampled, the within-response continuation indicator is measurable after the current action, and the recurrence resets at the response boundary. Arbitrary future-dependent masks and outcome-dependent response filtering require a separate argument.

At $c = 0$, $Q_t^c = Q_t^\mu$. At $c = \rho$, every switch decision at a target-supported pair succeeds, since $\lambda_t = 1$ there, so all later actions follow π and $Q_t^c = Q_t^\pi$; at a target-null pair $\rho_t(x, a) = 0$, and the convention $\lambda_t := 0$ instead makes the process revert to μ at once, which is immaterial because every such pair enters the identities above multiplied by $\rho_t(x, a) = 0$. These correspond to $\bar{c} = 0$ and to any $\bar{c}$ at least the supremum of ρ_t over behavior-supported pairs — equivalently, to the limit $\bar{c} \to \infty$ — for the adaptive threshold family. Intermediate traces mix the two continuation laws, but without further assumptions this does not imply a pointwise ordering such as $Q_t^\mu \le Q_t^c \le Q_t^\pi$.

C.1 The state-value V-trace–Retrace operator triangle

The switching target also has an exact operator interpretation. Define its behavior-action marginal directly by

$$[\mathcal{V}_c^{\mathrm{s}} V]_t(x) := \sum_{a \in \mathcal{A}_\mu(x)} \mu(a \mid x)[\mathcal{V}_c^{\mathrm{sa}} V]_t(x, a). \tag{27}$$

Set $[\mathcal{V}_c^{\mathrm{s}} V]_T = 0$. Appendix G shows that $\mathcal{V}_c^{\mathrm{s}} : \mathbb{V} \to \mathbb{V}$ is precisely state-value V-trace when its temporal-difference residual uses the exact ratio ρ_t; canonical IMPALA V-trace may cap that residual ratio separately (Espeholt et al., 2018). To define the corresponding action-value operator, let

$$[\mathcal{T}^\pi Q]_t(x, a) := \begin{cases} \mathbb{E}\left[\sum_{a' \in \mathcal{A}_\mu(x_{t+1})} \pi(a' \mid x_{t+1}) Q_{t+1}(x_{t+1}, a') \,\middle|\, x_t = x, a_t = a\right], & t < T - 1, \\ Q_{T-1}^\mu(x, a), & t = T - 1, \end{cases}$$

Thus $\mathcal{T}^\pi : \mathbb{Q} \to \mathbb{Q}$. (No reward term appears because every nonterminal reward is zero here; the terminal reward R enters only through the boundary value Q_{T-1}^μ.) Define the next-action-indexed finite-horizon Retrace-form operator $\mathcal{R}_c^\pi : \mathbb{Q} \to \mathbb{Q}$ directly by

$$[\mathcal{R}_c^\pi Q]_{T-1} = [\mathcal{T}^\pi Q]_{T-1}, \tag{28}$$

$$\begin{aligned} [\mathcal{R}_c^\pi Q]_t(x, a) = {} & [\mathcal{T}^\pi Q]_t(x, a) \\ & + \mathbb{E}_\mu\left[c_{t+1}\big([\mathcal{R}_c^\pi Q]_{t+1} - Q_{t+1}\big)(x_{t+1}, a_{t+1}) \mid x_t = x, a_t = a\right], \qquad t < T - 1. \end{aligned} \tag{29}$$

The trace coefficient is indexed by the next sampled action, as in canonical Retrace (Munos et al., 2016); the canonical choice is $c_j = \lambda \min\{1, \rho_j\}$. Equation (29) at $t = T - 2$ refers to c_{T-1}, which Definition B.1 never forms, since its products stop at c_{T-2}; we adopt the convention $c_{T-1} := 0$. Nothing in this paper depends on the choice, because every use takes $Q = \mathcal{B}V$. For such an input the terminal cases of the two definitions give $[\mathcal{B}V]_{T-1} = Q_{T-1}^\mu$ and $[\mathcal{R}_c^\pi(\mathcal{B}V)]_{T-1} = [\mathcal{T}^\pi(\mathcal{B}V)]_{T-1} = Q_{T-1}^\mu$ for every $V \in \mathbb{V}$, so $[\mathcal{R}_c^\pi Q]_{T-1} - Q_{T-1} = 0$ and the coefficient multiplies zero.

Theorem C.4 (State-action V-trace–Retrace decomposition) *Assume Equation* (13)*, so that* $\lambda_t = c_t/\rho_t$ *lies in* $[0, 1]$ *with the stated zero convention. For every* $V \in \mathbb{V}$*,*

$$\mathcal{B}(\mathcal{V}_c^{\mathrm{s}} V) = \mathcal{R}_c^\pi(\mathcal{B}V), \tag{30}$$

and, coordinatewise,

$$\begin{aligned} [\mathcal{V}_c^{\mathrm{sa}} V]_t &= (\rho_t - c_t)[\mathcal{B}V]_t + c_t[\mathcal{R}_c^\pi(\mathcal{B}V)]_t \\ &= \rho_t\left((1 - \lambda_t)[\mathcal{B}V]_t + \lambda_t[\mathcal{R}_c^\pi(\mathcal{B}V)]_t\right), \end{aligned}$$

where the bracket in the second line is a convex combination of the two targets because $\lambda_t \in [0, 1]$.

At $V = V^\mu$, $\mathcal{B}V^\mu = Q^\mu$ and

$$\mathbb{E}_\mu[R\widetilde{\rho}_t \mid x_t, a_t] = [\mathcal{V}_c^{\mathrm{sa}} V^\mu]_t(x_t, a_t) = \rho_t \left[(1-\lambda_t)Q_t^\mu + \lambda_t[\mathcal{R}_c^\pi Q^\mu]_t\right](x_t, a_t), \tag{31}$$

which is exactly the switching identity in Theorem C.1.

Three operators, three domains. State-value V-trace maps $\mathbb{V} \to \mathbb{V}$; state-action V-trace maps $\mathbb{V} \to \mathbb{Q}$ and directly supplies the target needed by policy optimization; Retrace maps $\mathbb{Q} \to \mathbb{Q}$. Marginalizing the current behavior action gives the first map, while the Bellman lift in Equation (30) connects it to the third. These are linked targets, not identical actor updates.

Appendix G gives the residual-sum forms and proofs.

This critic-free realization is specific, not a claim that abstract state-value V-trace or Retrace intrinsically require learned critics. Naively substituting the same terminal return into every classical temporal-difference residual is invalid—for example, it makes a nonterminal state-value residual $\rho_s(R - R)$ vanish. Separately derived complete-trajectory estimators do exist at the special inputs V^μ and Q^μ; see Appendix J.1. The benefit of the $\mathbb{V} \to \mathbb{Q}$ construction is that it directly yields the action-conditioned target needed by policy optimization and has the simple sample realization $R\widetilde{\rho}_t$.

The target Q_t^c itself depends on the learner through future actions and, for many rules, through the traces. Equation (22) therefore differentiates the current action probability while holding the continuation target fixed. Its precise local scalar interpretation is an anchored semigradient. At an anchor $\bar{\pi} = \pi_{\bar{\theta}}$, construct trace coefficients $c^{\bar{\pi}}$ and switching values $Q_t^{c^{\bar{\pi}}}$ from $(\bar{\pi}, \mu)$, and hold the latter fixed in

$$\mathcal{J}^\mu_{c^{\bar{\pi}},\bar{\pi}}(\theta) := \sum_{t=0}^{T-1} \mathbb{E}_{x_t \sim d_t^\mu}\left[\sum_{a \in \mathcal{A}_\mu(x_t)} \pi_\theta(a \mid x_t) Q_t^{c^{\bar{\pi}}}(x_t, a)\right]. \tag{32}$$

Then

$$\nabla_\theta \mathcal{J}^\mu_{c^{\bar{\pi}},\bar{\pi}}(\theta)\Big|_{\theta=\bar{\theta}} = G_{c^{\bar{\pi}}}(\bar{\pi}; \mu). \tag{33}$$

At $c^{\bar{\pi}} = 0$ this is the fixed-behavior Q^μ surrogate. At $c_t^{\bar{\pi}} = \rho_t^{\bar{\pi}}$, where $\rho_t^{\bar{\pi}} := \bar{\pi}(a_t \mid x_t)/\mu(a_t \mid x_t)$ is the anchor-to-behavior action ratio, it is the frozen-$Q^{\bar{\pi}}$ full-suffix semigradient. The two endpoints are the cases $\lambda_t \equiv 0$ and $\lambda_t \equiv 1$ of Equation (21) read at the anchor (Appendix F). The second is the less immediate of the two: at $\lambda \equiv 1$ the correction term in Equation (29) cancels the input, leaving $[\mathcal{R}_c^\pi Q]_t = [\mathcal{T}^\pi(\mathcal{R}_c^\pi Q)]_t$ for $t < T-1$ with the boundary $[\mathcal{R}_c^\pi Q]_{T-1} = Q_{T-1}^\mu$, which is the target-policy Bellman recursion; its unique finite-horizon solution is $Q^{\bar{\pi}}$ for every input Q. This does not assert a self-consistent global objective whose full derivative equals the V-TRACE PO direction throughout training.

The theorem characterizes the unclipped direction. Turning that direction into an LLM training objective requires an autodifferentiable ratio and an explicit decision about where clipping is applied.

D FROM THE TRACE ESTIMATOR TO A PRACTICAL OPTIMIZER

D.1 CLIPPING-PLACEMENT VARIANTS

The same unclipped trace direction can produce different optimizers once PPO clipping is introduced. V-TRACE PO, the method of the main text, clips the current-token ratio and leaves a detached continuation multiplier outside that mask; V-TRACE PO-TC instead clips the transformed trace. The primary comparison of the two placements is the radius-by-placement factorial of Section 6.1; Appendix L.3 reports the superseded single-radius study at $\bar{c} = 1$, $\epsilon = .2$. V-TRACE PG, which removes PPO clipping, remains a variance and numerical-stability diagnostic. Here b indexes responses within the current GPU microbatch, $\mathrm{sg}(u)$ denotes a stop-gradient copy of u, $\mathrm{clip}(u, l, h) = \min\{h, \max\{l, u\}\}$, and $M_{bt} \in \{0, 1\}$ marks valid response tokens in a padded batch.

A shared transformed ratio. At each forward evaluation, compute the current ratio ρ_t and suffix trace $\widetilde{\rho}_t$, then detach the trace-to-local-ratio multiplier. For $\rho_t > 0$, define the transformed ratio

$$\hat{\rho}_t := \mathrm{sg}\left(\frac{\widetilde{\rho}_t}{\rho_t}\right)\rho_t. \tag{34}$$

Table 3: Optimization variants. All three share the same per-token ascent direction when no PPO clip is active. The two clipped variants differ in which quantity determines the PPO derivative mask; the unclipped variant is a stress diagnostic.

Name	PPO-clipped quantity	Purpose
V-TRACE PO-TC	Transformed ratio $\hat{\rho}_t$	Suffix trace itself determines the PPO derivative mask
V-TRACE PO	Local ratio ρ_t	Detached suffix multiplier lies outside the local PPO derivative mask
V-TRACE PG	None	Unclipped variance and numerical-stability stress test

Its forward value and autodifferentiated gradient are

$$\hat{\rho}_t = \widetilde{\rho}_t, \qquad \nabla_\theta \hat{\rho}_t = \widetilde{\rho}_t \nabla_\theta \log \pi_\theta(a_t \mid x_t). \tag{35}$$

A division-free implementation is preferable in finite precision. With $\delta_t(\theta) = \log \pi_\theta(a_t \mid x_t) - \log \mu(a_t \mid x_t)$, use

$$\hat{\rho}_t = \mathrm{sg}(\widetilde{\rho}_t) \exp\left(\delta_t(\theta) - \mathrm{sg}(\delta_t(\theta))\right). \tag{36}$$

Equations (34) and (36) have the same forward value and local autodifferentiated gradient, but the latter never divides by a possibly underflowed ρ_t and is therefore the recommended implementation of the transformed ratio. The detached multiplier of V-TRACE PO avoids the division separately, through $w_t = 1 + \lambda_t(\widetilde{\rho}_{t+1} - 1)$ below. Note that the $[-20, 20]$ clamp bounds $\rho_t \geq e^{-20} \approx 2.06 \times 10^{-9}$, which is safe in bfloat16 and float32 but underflows to zero in float16. An actually sampled action with nonfinite behavior log-probability indicates invalid metadata. A mathematically target-null action contributes zero to the unnormalized trace but has no log score; it must be handled by an explicit mask, not an ε-modified denominator presented as exact.

V-TRACE PO-TC: transformed-trace PPO. Given the detached response-level advantage A_b — the mean-only group advantage $R_b - \bar{R}$ used by every reported experiment from the radius-by-placement factorial onward, with the standard-deviation-normalized Equation (56) as the documented exception — the transformed-trace implementation uses the PPO-style objective

$$\mathcal{J}_{\mathrm{trace}} = \frac{1}{\sum_{b,t} M_{bt}} \sum_{b,t} M_{bt} \min\left\{\hat{\rho}_{bt} A_b, \mathrm{clip}(\hat{\rho}_{bt}, 1-\epsilon, 1+\epsilon) A_b\right\}. \tag{37}$$

The quantity entering its clipping rule has the forward value of the complete suffix trace, $\hat{\rho}_{bt} = \widetilde{\rho}_{bt}$. Because the experiments substitute a group-relative A_b for the return R_b, this is a practical group-centered surrogate; Corollary C.3 applies exactly under the mean-only convention and not under standardization. Each objective is a valid-token mean within one microbatch; gradient accumulation averages the resulting microbatch gradients. The switching theorem characterizes its return-weighted, unclipped core, not the complete group-normalized PPO update.

V-TRACE PO: local-ratio clipping. For $\rho_{bt} > 0$, define the nonnegative detached continuation multiplier

$$w_{bt} := \mathrm{sg}\left(\frac{\widetilde{\rho}_{bt}}{\rho_{bt}}\right).$$

Equivalently, writing τ_b for response b's final token index, $w_{b,\tau_b} = 1$ and, before termination, $w_{bt} = \mathrm{sg}(1 + \lambda_{bt}(\widetilde{\rho}_{b,t+1} - 1))$, where $\lambda_{bt} = c_{bt}/\rho_{bt}$. The implemented alternative clips only the differentiable current-token ratio and places the continuation multiplier outside the clipping operation:

$$\mathcal{J}_{\mathrm{local}} = \frac{1}{\sum_{b,t} M_{bt}} \sum_{b,t} M_{bt}\, w_{bt} \min\left\{\rho_{bt}(\theta) A_b, \mathrm{clip}(\rho_{bt}(\theta), 1-\epsilon, 1+\epsilon) A_b\right\}. \tag{38}$$

Whenever the local PPO clip is inactive, its per-token ascent direction is $A_b \widetilde{\rho}_{bt} s_{bt}$. When clipping is active, however, the derivative mask depends on ρ_{bt} rather than on the cumulative trace. This defines a practical optimizer, but its clipping rule is not another exact consequence of the switching theorem.

V-TRACE PG: no PPO clipping. A third objective removes PPO clipping and exposes the detached trace weight directly:

$$\mathcal{J}_{\mathrm{VPG}} = \frac{1}{\sum_{b,t} M_{bt}} \sum_{b,t} M_{bt} A_b \, \hat{\rho}_{bt}. \tag{39}$$

Its per-token ascent direction is $A_b \widetilde{\rho}_{bt} s_{bt}$. It coincides with the sampled direction in Equation (15) when $A_b = R_b$, up to the valid-token normalization; group-relative or otherwise transformed advantages define an additional practical modification. This objective is the cleanest estimator diagnostic, but also the least protected from heavy-tailed suffix weights.

Endpoints and clipping placement. For the threshold family $c_t = \min\{\bar{c}, \rho_t\}$, $\bar{c} = 0$ gives $\widetilde{\rho}_t = \rho_t$ and $w_t = 1$. Consequently, Equations (37) and (38) both reduce exactly to ordinary current-token PPO/GRPO-style clipping, provided the advantage, behavior denominator, and token-versus-response reduction are matched. This is an exact-arithmetic statement. Our implementation matches a local baseline with the same straight-through safeguard: the trace's forward ratio uses the detached, clipped log-ratio, while the derivative uses the raw learner score. It does not match a baseline that differentiates through the clamp, and it is not an exact change of measure on tokens where the clamp activates. Equation (39) instead reduces to unclipped current-token importance sampling. If the data are exactly on-policy, $\rho_t = 1$ everywhere and all three objectives have the same per-token gradient because both PPO clips are inactive.

The setting $\bar{c} = 1$ uses $c_t = \min\{1, \rho_t\}$; it is a clipped, state–action-dependent suffix correction, not the constant-$\lambda = 1$ trace. Taking the formal endpoint $\bar{c} = \infty$ gives $c_t = \rho_t$, $\widetilde{\rho}_t = \rho_{t:T-1}$, and $w_t = \rho_{t+1:T-1}$. This is full-suffix correction while the sampled prefix remains under d_t^{μ}; it is not full-trajectory importance sampling. Completed transformed-trace experiments and pilot diagnostics approximate this endpoint with the finite cap $\bar{c} = 10^7$ and report it as a stress condition. The matched local-clipping study excludes it because the full future-ratio product w_t would remain outside the PPO mask.

Other token-versus-response reductions must be stated explicitly because they change length weighting. Proposition C.2 and Corollary C.3 give a sharper statement for the unclipped exact-ratio core: a detached mean-only RLOO reward baseline preserves its expected direction, and inclusive group-mean subtraction differs only by the known factor $(K-1)/K$. Dividing by a realized group standard deviation is not justified by that control-variate argument because the denominator depends on the sampled reward of the same trajectory. Group standardization, other reward normalization, KL terms, and entropy terms are therefore additional practical transformations rather than parts of the exact target identity.

Most importantly, PPO clipping introduces an advantage-sign-dependent derivative mask. Its derivative may be zero when Equation (35) is nonzero. Global gradient clipping then transforms the aggregate update again. We therefore make no claim that Equation (37) preserves the switching target, defines a trust region, or inherits a monotone-improvement guarantee. The placement of the PPO clip is especially consequential at the large-cap endpoint. In Equation (37), the full suffix product itself is compared with $1 \pm \epsilon$; large positive-advantage upper tails and large negative-advantage lower tails can therefore receive zero policy gradient. Equation (38) clips only the local ratio, leaving the detached suffix product as a potentially very large multiplier, while Equation (39) has no PPO derivative mask. All three may still be changed by optimizer-level gradient clipping. Thus similar empirical behavior for the $\bar{c} = 10^7$ large-cap proxy under transformed-trace clipping could reflect PPO saturation rather than similar raw V-trace estimators. Section 6.1 reports the direct comparison at the controlled cap $\bar{c} = 1$.

D.2 IMPLEMENTATION IN ONE REVERSE SCAN

Algorithm 1 turns the estimator into a masked batch operation and exposes the entire change needed on top of a conventional PPO/GRPO loss. The inputs are differentiable learner log-probabilities, fixed stored behavior log-probabilities, detached group advantages, and two masks. Let $M \in \{0,1\}^{B \times T}$ mark valid response tokens and let $e_{bt} = 1$ exactly when the edge $t \to t+1$ stays inside the same nonterminal response. For the un-packed batches in our experiments, $e_{bt} = M_{bt} M_{b,t+1}$; packed implementations must additionally require the two tokens to share a segment identifier. This rule resets the recurrence at a length-capped response as though it had terminated, which is the finite-

horizon convention discussed below rather than a claim that the truncation is a terminal boundary; we report cap-hit rates accordingly.

Algorithm 1: Drop-in V-TRACE PO loss for PPO/GRPO, with the V-TRACE PO-TC placement and the unclipped V-TRACE PG diagnostic as non-default options

```
import math
import torch

def suffix_trace(rho, continues, c_bar):
    """continues[:, t] is 1 only across a valid t -> t+1 edge."""
    assert c_bar >= 0
    c = rho.clamp(max=c_bar)
    trace = rho.clone()
    for t in reversed(range(rho.shape[1] - 1)):
        trace[:, t] = rho[:, t] + continues[:, t] * c[:, t] * (
            trace[:, t + 1] - 1
        )
    return trace

def continuation_weight(log_rho, trace, continues, c_bar):
    """w_t = 1 + lambda_t (trace_{t+1} - 1), with lambda_t = c_t / rho_t.

    Exactly equal to trace_t / rho_t, but never divides: for the
    threshold
    family lambda_t = min(1, c_bar / rho_t) = exp(min(0, log c_bar - log
    rho)).
    """
    if c_bar == 0:
        return torch.ones_like(trace)
    lam = (math.log(c_bar) - log_rho).clamp(max=0.0).exp()
    nxt = torch.cat([trace[:, 1:], torch.ones_like(trace[:, :1])], dim=1)
    return 1.0 + continues * lam * (nxt.detach() - 1.0)

def ppo_token_loss(ratio, advantage, epsilon):
    unclipped = -advantage * ratio      # both are (B, T)
    clipped = -advantage * ratio.clamp(1 - epsilon, 1 + epsilon)
    return torch.maximum(unclipped, clipped)


def vtrace_po_loss(logp, behavior_logp, advantage, token_mask,
                   continues, c_bar=1.0, epsilon=0.2,
                   clip_placement="local"):
    """All tensors are (B, T); `advantage` may be (B,) response-level."""
    # Forward-only IS values; [-20, 20] is a numerical safeguard.  It
    bounds
    # rho >= e^-20 ~ 2.06e-9: safe in fp32/bf16, flushes to zero in fp16.
    log_rho = (logp - behavior_logp.detach()).detach().clamp(-20, 20)
    valid = token_mask.bool()
    rho = torch.where(valid, log_rho.exp(), torch.ones_like(log_rho))
    trace = suffix_trace(rho, continues, c_bar)

    # Both surrogates retain the current learner score derivative.  The
    # difference is masked first: padded positions may carry -inf, and
    # (-inf) - (-inf) = NaN would survive the 0 * NaN in the reduction
    below.
    score_delta = torch.where(valid, logp - logp.detach(),
                              torch.zeros_like(logp))
    score_factor = torch.exp(score_delta)
    local_ratio = rho.detach() * score_factor
    trace_ratio = trace.detach() * score_factor
    advantage = advantage.detach()
    if advantage.dim() == 1:            # response-level -> broadcast on
    tokens
```

```
        advantage = advantage.unsqueeze(-1)

    if clip_placement == "local":       # V-Trace PO (default)
        continuation = continuation_weight(
            log_rho, trace.detach(), continues, c_bar
        )
        continuation = torch.where(valid, continuation,
                                   torch.zeros_like(continuation))
        token_loss = continuation * ppo_token_loss(
            local_ratio, advantage, epsilon
        )
    elif clip_placement == "trace":     # V-Trace PO-tc
        token_loss = ppo_token_loss(trace_ratio, advantage, epsilon)
    elif clip_placement == "none":      # V-Trace PG (diagnostic,
    unclipped)
        token_loss = -advantage * trace_ratio
    else:
        raise ValueError(f"unknown clip placement: {clip_placement}")

    assert torch.isfinite(token_loss[valid]).all()
    token_loss = torch.where(valid, token_loss, torch.zeros_like(
    token_loss))
    return (token_loss * token_mask).sum() / token_mask.sum().clamp(min
    =1)
```

All three branches reuse the same reverse scan. The scan is linear in response length, uses no learned value head, and is recomputed after each learner update while the stored behavior log-probabilities remain fixed. The log-ratio clamp is solely a finite-precision safeguard; omitting it recovers the exact mathematical ratio. The score-factor attachments give the displayed forward ratios while retaining their local score derivatives. In the default V-TRACE PO branch, the continuation weight is formed from $w_t = 1 + \lambda_t(\widetilde{\rho}_{t+1} - 1)$, which equals $\widetilde{\rho}_t/\rho_t$ exactly while never dividing by the ratio, and is detached before it scales the token loss. At $\bar{c} = 0$, it equals one and the branch reduces exactly to the matched PPO/GRPO implementation.

Correctness still depends on the provenance of behavior probabilities, terminal and packing masks, and reduction semantics. In particular, a token cap that truncates a live trajectory is not a terminal boundary: it requires rejection, an explicit truncated objective, or a boundary estimator. Appendix M gives the complete experimental and logging protocol.

E RELATED WORK: WHAT IS BEING CORRECTED?

The closest methods differ less in whether they are "off-policy" than in *which part* of the mismatch they control: state visitation, the current action, the future continuation, or one aggregate response score.

Off-policy policy gradients and state weighting. Classical off-policy actor–critic methods distinguish the action ratio from the state-distribution weighting induced by the behavior policy (Degris et al., 2012; Imani et al., 2018). Emphatic methods correct the latter through follow-on state weights; our estimator deliberately retains d_t^{μ} and instead changes the future continuation attached to the current score. This is why the natural interpretation is the anchored semigradient in Equation (33), not the exact gradient of a generic on-policy objective.

V-trace, Retrace, and actor–critic replay. Per-decision importance sampling and eligibility traces provide the classical multi-step lineage (Precup et al., 2000). IMPALA uses state-value V-trace inside an asynchronous actor–learner architecture: its actor update multiplies a clipped current-action ratio by an advantage bootstrapped from a learned V-trace value target (Espeholt et al., 2018). Retrace is an action-value evaluation operator with next-action-indexed traces (Munos et al., 2016); ACER combines a Retrace critic with truncated actor ratios, explicit bias correction, and a trust-region projection (Wang et al., 2017).

Our construction is not a relabeling of any of these actor updates. It maps a state value to an action-conditioned target and, on complete terminal-reward trajectories, collapses directly to a return-weighted suffix trace. Its behavior-action marginal recovers exact-ratio state-value V-trace, while its backed-up marginal follows the Retrace recursion. The exact operator triangle is defined in Section C.1; at V^μ, Equation (31) identifies $[\mathcal{R}_c^\pi Q^\mu]_t$ as the forced-next-target-action branch, while the complete switching value Q_t^c mixes it once with Q_t^μ. This is an affine target identity, not equality to canonical IMPALA or ACER actor updates. Canonical V-trace defines separate truncations for the residual ratio and trace coefficient; when residual-ratio truncation binds, its fixed point is the value of a clipped intermediate policy rather than V^π. Appendix G gives the residual forms and proofs.

Multi-step off-policy learning. $Q(\lambda)$ variants, adaptive traces, Taylor policy optimization, and DoMo-AC also study multi-step correction and its bias–variance trade-off (Harutyunyan et al., 2016; Rowland et al., 2020; Kozuno et al., 2021; Tang et al., 2020; 2023). We treat these as related constructions; we do not call them special cases without a sample-wise or expectation-level reduction. Shared-experience actor–critic work provides a complementary replay perspective (Schmitt et al., 2020).

LLM policy surrogates. PPO clips a local likelihood-ratio surrogate (Schulman et al., 2017). GRPO and RLOO supply critic-free group or leave-one-out baselines (Shao et al., 2024; Ahmadian et al., 2024); DAPO and Dr. GRPO modify the sampling, normalization, and token-versus-response reduction around that local update (Yu et al., 2025; Liu et al., 2025). These choices must be held fixed in a trace comparison because they change length weighting and gradient scale.

Among concurrent LLM methods, NFPO is closest in mechanism (Yoon et al., 2026). It multiplies the current-token ratio by a detached, clipped product of the next $N-1$ ratios and applies a separate token-level trust-region mask. Its fixed horizon truncates correction by distance; V-TRACE PO instead uses a centered recurrence with per-decision coefficients, which induces the switching target and V-trace–Retrace relations above. Because NFPO's future product multiplies a separately masked local ratio, its clipping placement is closer to V-TRACE PO than to the original V-TRACE PO-TC objective. As in V-TRACE PO, NFPO's detached future multiplier yields an anchored semigradient; its performance-difference bound does not directly characterize the complete practical optimizer. A matched implementation comparison remains important.

GSPO assigns a length-normalized geometric-mean ratio to an entire response (Zheng et al., 2025a). PNPO instead compresses the cumulative *prefix* ratio at each position and gates updates using that prefix statistic (Zhang et al., 2026a). CTPO starts from the cumulative prefix ratio and uses position-dependent log-space clipping, while MinPRO retains ρ_t but replaces the preceding-prefix product by the minimum preceding local ratio (Zhang et al., 2026b; Lei et al., 2026). These methods and V-TRACE PO operate on complementary temporal directions: the former use an exact cumulative ratio or compressed causal-prefix surrogates for accumulated mismatch, whereas V-TRACE PO retains behavior-distributed prefixes and corrects the suffix continuation. None is identical to full-trajectory importance sampling.

Recent stale-rollout methods primarily redesign clipping or masking. CISPO caps a detached local importance coefficient while retaining token gradients (MiniMax, 2025); DPPO replaces sampled-ratio clipping by masks based on estimated policy divergence (Qi et al., 2026); M2PO masks extreme tokens under a batch constraint on squared sampled log ratios (Zheng et al., 2025b); BAPO adapts asymmetric PPO bounds to rebalance positive and negative updates (Xi et al., 2025); and Mu-GRPO combines a relaxed local clip with a negative-advantage veto under aggressive rollout reuse (Tian et al., 2026). VCPO instead responds to collapsing effective sample size with an adaptive learning rate and an off-policy minimum-variance baseline (Huang et al., 2026). These methods reshape the magnitude, clipping bounds, admission, or aggregate step size of local-token updates. V-TRACE PO changes the continuation each action is scored against, so the approaches are complementary as well as direct empirical comparators.

Off-policy data and asynchronous systems. Two lines of work make the treatment of positive and negative examples deliberately asymmetric, by different means. Asymmetric REINFORCE does it through the baseline, taking $A = r - V$ with V tunable, and shows a policy-improvement guarantee when V lower-bounds the expected reward (Arnal et al., 2025); tapered off-policy REINFORCE instead applies an asymmetric, tapered importance weight (Le Roux et al., 2025). Soft Policy

Table 4: Where representative methods intervene. Rows summarize mechanisms rather than performance.

Method	Temporal object	Critic	Main control or distinction
PPO / GRPO / Dr. GRPO	Current token	Varies	Local ratio clipping; $\bar{c} = 0$ endpoint after matching the advantage and reduction
GSPO	Response ratio	No	Length-normalized geometric mean; no time-dependent suffix target
CTPO / MinPRO / PNPO	Causal prefix	No	CTPO uses the exact cumulative ratio; MinPRO/PNPO compress preceding ratios
NFPO	Fixed suffix window	No	Clipped future product outside a separate token mask
M2PO	Token + batch statistic	No	Squared-log-ratio trust mask; no suffix change of measure
IMPALA / V-trace	State-value TD trace	Yes	Trace and residual-ratio truncation; exact-ratio marginal recovered here
ACER / Retrace	Action-value trace	Yes‡	Truncated ratios, bias correction, and trust-region projection
V-TRACE PO-TC	Adaptive suffix	No†	Coefficient-capped trace itself determines the PPO mask
V-TRACE PO	Adaptive suffix	No†	Same trace direction; local ratio determines the PPO mask

†For complete terminal-reward trajectories. ‡As originally posed. On complete terminal-reward trajectories Retrace likewise admits a reward-only estimator (Appendix J.1), so the critic column separates the methods as published rather than as an intrinsic property of the operators.

Optimization and AGRO optimize broader mixtures of on- and off-policy data under different consistency objectives (Cohen et al., 2025; Tang et al., 2025), while LUFFY mixes teacher traces with on-policy rollouts (Yan et al., 2025). When behavior probabilities are unavailable, such data are outside our exact correction; compatible tokenization is required as well. At the systems level, LlamaRL and AReaL expose natural actor–learner lag (Wu et al., 2025; Fu et al., 2025). LlamaRL's AIPO upper-clips local importance ratios, while AReaL's decoupled PPO separates behavior-policy correction from a more recent proximal anchor; both remain local in time. Numerical differences between generation and training can create mismatch even at the same nominal checkpoint (Yao et al., 2025; He & Thinking Machines Lab, 2025).

Table 4 summarizes the main structural distinctions.

The distinctive contribution is the centered, coefficient-truncated suffix recurrence, its switching and operator analysis, and its realization inside PPO/GRPO. V-TRACE PO-TC and V-TRACE PO share that construction but expose it to different derivative masks. Prefix weighting, adaptive clipping, proximal anchoring, and data selection solve different pieces of the off-policy problem and can in principle be combined with either placement. The matched-clipping study below isolates suffix correction under local clipping and compares the two placements directly.

This appendix supplies the finite-horizon proofs and operator details behind the main claims, and Appendix M gives the protocols for the experiments, seed-level results, sensitivity analyses, mechanism studies, and validation criteria. The preceding appendix develops the theory and experiments at greater length; where the two overlap, the statements proved here are canonical. In the proofs, $\mathbb{E}_\mu$ denotes expectation over trajectories generated by the behavior policy μ, while an unsubscripted $\mathbb{E}$ denotes expectation over the environment transition alone, conditional on the stated (x_t, a_t), and therefore does not depend on any policy.

F CHANGE-OF-MEASURE AND TRACE PROOFS

F.1 PROOF OF PROPOSITION 2.1

Write $s(x, a) := \nabla_\theta \log \pi_\theta(a \mid x)$, so that $s_t = s(x_t, a_t)$; $\mathbb{E}_\mu$ is the expectation over trajectories generated by μ from the common initial law, and all statements are read on the behavior-supported set $\mathcal{Z}_t^\mu$. Three change-of-measure steps suffice.

(a) Prefix reweighting. For any function f of the prefix state,

$$\mathbb{E}_\mu \left[\rho_{0:t-1} f(x_t)\right] = \mathbb{E}_{x_t \sim d_t^\pi} \left[f(x_t)\right]. \tag{40}$$

A prefix law is the initial law times the action probabilities times the transition kernels; kernels and initial law are policy independent, so multiplying the μ-law of $(x_0, a_0, \ldots, x_t)$ by $\rho_{0:t-1}$ returns its π-law, and summing over prefixes ending at x_t gives d_t^π. Coverage enters here: no prefix of positive π-probability has zero μ-probability, so no π-mass is lost.

(b) Current-action reweighting. For any state x and function g,

$$\mathbb{E}_{a\sim\mu(\cdot|x)}\left[\frac{\pi(a\mid x)}{\mu(a\mid x)}g(x,a)\right]=\sum_{a\in\mathcal{A}_\mu(x)}\pi(a\mid x)g(x,a)=\mathbb{E}_{a\sim\pi(\cdot|x)}\left[g(x,a)\right], \tag{41}$$

the second equality by coverage, that is, $\pi(a\mid x)=0$ off $\mathcal{A}_\mu(x)$.

(c) Suffix reweighting. For $(x,a)\in\mathcal{Z}_t^\mu$,

$$\mathbb{E}_\mu\left[R\,|\,x_t=x,a_t=a\right]=Q_t^\mu(x,a),\qquad \mathbb{E}_\mu\left[R\,\rho_{t+1:T-1}\,|\,x_t=x,a_t=a\right]=Q_t^\pi(x,a); \tag{42}$$

the first is the definition of the restart value, the second applies (a) to the continuation, whose law depends on the past only through (x_t,a_t).

Let $h_t^\nu(x):=\mathbb{E}_{a\sim\pi(\cdot|x)}[Q_t^\nu(x,a)s(x,a)]$, so that $\mathbb{E}_{x_t\sim d_t^{\nu'}}[h_t^\nu(x_t)]$ is exactly a right-hand side of the proposition, and abbreviate $Q_t^\nu:=Q_t^\nu(x_t,a_t)$, $h_t^\nu:=h_t^\nu(x_t)$. Conditioning first on (x_t,a_t), under which $\rho_{0:t}$ and s_t are measurable and factor out, and then on x_t, the tower property gives

$$\begin{aligned}
&(3)\ \ \mathbb{E}_\mu[R\,\rho_{0:T-1}s_t]\overset{(c)}{=}\mathbb{E}_\mu[\rho_{0:t-1}\rho_tQ_t^\pi s_t]\overset{(b)}{=}\mathbb{E}_\mu[\rho_{0:t-1}h_t^\pi]\overset{(a)}{=}\mathbb{E}_{x_t\sim d_t^\pi}[h_t^\pi],\\
&(4)\ \ \mathbb{E}_\mu[R\,\rho_{t:T-1}s_t]\overset{(c)}{=}\mathbb{E}_\mu[\rho_tQ_t^\pi s_t]\overset{(b)}{=}\mathbb{E}_\mu[h_t^\pi]=\mathbb{E}_{x_t\sim d_t^\mu}[h_t^\pi],\\
&(5)\ \ \mathbb{E}_\mu[R\,\rho_{0:t}s_t]\overset{(c)}{=}\mathbb{E}_\mu[\rho_{0:t-1}\rho_tQ_t^\mu s_t]\overset{(b)}{=}\mathbb{E}_\mu[\rho_{0:t-1}h_t^\mu]\overset{(a)}{=}\mathbb{E}_{x_t\sim d_t^\pi}[h_t^\mu],\\
&(6)\ \ \mathbb{E}_\mu[R\,\rho_ts_t]\overset{(c)}{=}\mathbb{E}_\mu[\rho_tQ_t^\mu s_t]\overset{(b)}{=}\mathbb{E}_\mu[h_t^\mu]=\mathbb{E}_{x_t\sim d_t^\mu}[h_t^\mu].
\end{aligned}$$

Every row uses (42) to replace the return by a continuation value. The final equality in the second and fourth lines is the definition of d_t^μ; those rows never invoke (40), which is why they leave the prefix-state law at d_t^μ. The support assumptions are used only in (40) and (41), and every ratio above is finite on $\mathcal{Z}_t^\mu$.

The full-trajectory identity in Equation (10) follows by summing the first row over t. For the current-token surrogate, both d_t^μ and Q_t^μ are fixed during differentiation. Applying $\nabla_\theta\pi_\theta=\pi_\theta\nabla_\theta\log\pi_\theta$ to Equation (11), and then using the fourth row of Proposition 2.1, gives Equation (12).

The same argument proves the anchored identity in Equation (33): at the anchor, the state distribution and switching target are held fixed, so differentiation acts only on the current action probability. The endpoint interpretations follow from $Q^{c^{\bar\pi}}=Q^\mu$ at $c^{\bar\pi}=0$ and $Q^{c^{\bar\pi}}=Q^{\bar\pi}$ at $c_t^{\bar\pi}=\rho_t^{\bar\pi}$.

F.2 Trace recursion and endpoints

Separate the first summand $s=t+1$ in Equation (14) and factor c_t from the rest:

$$\begin{aligned}
\widetilde{\rho}_t&=\rho_t+c_t(\rho_{t+1}-1)+c_t\sum_{s=t+2}^{T-1}c_{t+1:s-1}(\rho_s-1)\\
&=\rho_t+c_t(\widetilde{\rho}_{t+1}-1).
\end{aligned}$$

For $c=0$ the recurrence stops immediately. For $c=\rho$, direct telescoping gives

$$\widetilde{\rho}_t=\rho_t+\sum_{s=t+1}^{T-1}(\rho_{t:s}-\rho_{t:s-1})=\rho_{t:T-1}.$$

Substituting $c_j=\lambda\rho_j$ and collecting each partial product yields Equation (17).

G State-action V-trace: equivalent forms and proofs

This section proves the operator results stated in Appendix B and C.1, using the spaces $\mathbb{V}$, $\mathbb{Q}$ and the residual δ_s^V defined there.

Proposition G.1 (Equivalent finite-horizon forms) *For every $V \in \mathbb{V}$ and $0 \le t < T$,*

$$[\mathcal{V}_c^{\mathrm{sa}} V]_t(x,a) = \mathbb{E}_\mu \Bigg[c_{t:T-2}\rho_{T-1} R + \sum_{s=t}^{T-2} c_{t:s-1}(\rho_s - c_s) V_{s+1}(x_{s+1}) \;\Big|\; x_t = x, a_t = a \Bigg]. \tag{43}$$

For $t = T-1$, the sum is empty and $c_{t:T-2} = 1$. The residual and telescoped forms are equivalent to the terminal boundary in Equation (19) *and the backward recursion in Equation* (20)*.*

Proof. In Equation (18), the exterior V_t cancels the first negative baseline. Each positive term $c_{t:s-1}\rho_s V_{s+1}$ then combines with the following negative term $-c_{t:s}V_{s+1}$, leaving $c_{t:s-1}(\rho_s - c_s)V_{s+1}$; the last positive term is $c_{t:T-2}\rho_{T-1}R$. This proves Equation (43). Separating the first residual in the defining sum and conditioning the remainder one step later gives Equation (20); its terminal case is Equation (19), and unrolling the recursion recovers the definition. □

Proposition G.2 (State-action marginal is state-value V-trace) *The behavior-action marginal $\mathcal{V}_c^{\mathrm{s}} V$ in Equation* (27) *has the standard exact-residual-ratio state-value V-trace form*

$$[\mathcal{V}_c^{\mathrm{s}} V]_t(x) = V_t(x) + \mathbb{E}_\mu \Bigg[\sum_{s=t}^{T-2} c_{t:s-1}\rho_s (V_{s+1}(x_{s+1}) - V_s(x_s)) + c_{t:T-2}\rho_{T-1}(R - V_{T-1}(x_{T-1})) \;\Big|\; x_t = x \Bigg],$$

with $[\mathcal{V}_c^{\mathrm{s}} V]_T = 0$.

Proof. Average Equation (18) over the current behavior action. The two residuals differ by $(\rho_s - 1)V_s(x_s)$, carried by the coefficient $c_{t:s-1} = \prod_{j=t}^{s-1} c_j$. Because x_s is the full prefix, both that coefficient and $V_s(x_s)$ are functions of x_s, so conditioning on x_s takes them outside and the remaining factor has conditional mean zero, $\mathbb{E}_\mu[\rho_s - 1 \mid x_s] = 0$. This identity uses the exact residual ratio. Replacing it separately by a capped ratio generally breaks the zero-mean equality. □

The Retrace recursion in Equations (28)–(29) unrolls to

$$[\mathcal{R}_c^\pi Q]_t(x,a) = Q_t(x,a) + \mathbb{E}_\mu \left[\sum_{s=t}^{T-1} c_{t+1:s}([\mathcal{T}^\pi Q]_s - Q_s) \,\middle|\, x_t = x, a_t = a \right].$$

The current residual has empty coefficient $c_{t+1:t} = 1$; backward substitution recovers the main-text recursion. Canonical Retrace commonly uses $c_j = \lambda \min\{1, \rho_j\}$; Theorem C.4 only needs $0 \le c_j \le \rho_j$.

G.1 Proof of Theorem C.4

Proof. Equations (19) and (20) imply

$$[\mathcal{V}_c^{\mathrm{sa}} V]_t = (\rho_t - c_t)[\mathcal{B}V]_t + c_t[\mathcal{B}(\mathcal{V}_c^{\mathrm{s}} V)]_t.$$

Averaging over the current behavior action and using $\mu(a \mid x)\rho_t(x,a) = \pi(a \mid x)$ gives

$$[\mathcal{B}(\mathcal{V}_c^{\mathrm{s}} V)]_t = [\mathcal{T}^\pi(\mathcal{B}V)]_t + \mathbb{E}_\mu\big[c_{t+1}\big([\mathcal{B}(\mathcal{V}_c^{\mathrm{s}} V)]_{t+1} - [\mathcal{B}V]_{t+1}\big) \mid x_t, a_t\big]$$

with the Retrace terminal boundary. This is the Retrace recursion for $Q = \mathcal{B}V$, so finite backward induction gives $\mathcal{B}(\mathcal{V}_c^{\mathrm{s}} V) = \mathcal{R}_c^\pi(\mathcal{B}V)$. Substitution proves both displayed forms. □

Corollary G.3 (Behavior-value specialization) *At $V = V^\mu$, $\mathcal{B}V^\mu = Q^\mu$; the theorem and switching recursion give*

$$Q_t^c = (1 - \lambda_t)Q_t^\mu + \lambda_t[\mathcal{R}_c^\pi Q^\mu]_t, \tag{44}$$

$$[\mathcal{V}_c^{\mathrm{sa}} V^\mu]_t = \rho_t Q_t^c = \mathbb{E}_\mu[R\widetilde{\rho}_t \mid x_t, a_t]. \tag{45}$$

This proves Equation (31)*. The Retrace target is the forced-next-target-action branch inside Q_t^c, not the already-mixed switching value itself. Moreover,*

$$\boxed{\mathcal{B}(\mathcal{V}_c^{\mathrm{s}} V^\mu) = \mathcal{R}_c^\pi Q^\mu.} \tag{46}$$

Proof. The identity $\mathcal{B}V^\mu = Q^\mu$ is the behavior Bellman equation. The recursion for $\mathcal{R}_c^\pi Q^\mu$ in Equation (29) becomes

$$[\mathcal{R}_c^\pi Q^\mu]_{T-1} = Q_{T-1}^\mu,$$

$$[\mathcal{R}_c^\pi Q^\mu]_t(x,a) = \mathbb{E}\left[\sum_{a' \in \mathcal{A}_\mu(x_{t+1})} \pi(a' \mid x_{t+1}) \times \left\{(1-\lambda_{t+1})Q_{t+1}^\mu + \lambda_{t+1}[\mathcal{R}_c^\pi Q^\mu]_{t+1}\right\}(x_{t+1}, a') \,\middle|\, x_t = x, a_t = a\right].$$

Backward induction with Equation (21) identifies $[\mathcal{R}_c^\pi Q^\mu]_t$ with its inner, forced-next-target-action expectation and therefore gives Equation (44). The remaining identities follow from Theorem C.4 and Theorem C.1. Finally, Proposition G.2 establishes the standard exact-ratio form of the marginal $\mathcal{V}_c^{\mathrm{s}} V^\mu$, while Equation (30) gives $\mathcal{B}(\mathcal{V}_c^{\mathrm{s}} V^\mu) = \mathcal{R}_c^\pi Q^\mu$, proving Equation (46). □

The three constructions are therefore distinct. State-value V-trace maps state values to state values, state-action V-trace maps state values to action-conditioned targets, and Retrace maps action values to action values. Equation (46) connects their outputs without identifying their domains or their actor updates.

H Proofs of the terminal-return and switching results

H.1 Proof of Proposition B.6

By Proposition G.1, the action-conditioned operator has the telescoped form in Equation (43). For $V = V^\mu$, each $V_{s+1}^\mu(x_{s+1}) = \mathbb{E}_\mu[R \mid x_{s+1}]$. Its multiplying factor is measurable at that history, so the tower property permits the same-trajectory sample R. The coefficient of R is

$$\begin{aligned} c_{t:T-2}\rho_{T-1} + \sum_{s=t}^{T-2} c_{t:s-1}(\rho_s - c_s) &= \rho_t + \sum_{s=t+1}^{T-1} c_{t:s-1}(\rho_s - 1) \\ &= \widetilde{\rho}_t, \end{aligned}$$

where the negative trace products telescope. This proves both the pathwise collapse and conditional unbiasedness in Proposition B.6.

H.2 Proof of Theorem C.1

At $t = T-1$, $\widetilde{\rho}_{T-1} = \rho_{T-1}$, so the conditional expectation in Theorem C.1 is $\rho_{T-1}Q_{T-1}^\mu = \rho_{T-1}Q_{T-1}^c$. Suppose the result holds at $t+1$. Equation (16) gives

$$\mathbb{E}_\mu[R\widetilde{\rho}_t \mid x_t, a_t] = \rho_t Q_t^\mu + c_t\left(\mathbb{E}_\mu[R\widetilde{\rho}_{t+1} \mid x_t, a_t] - Q_t^\mu\right).$$

Condition first on (x_{t+1}, a_{t+1}), apply the induction hypothesis, and use $\mu(a' \mid x')\rho_{t+1}(x', a') = \pi(a' \mid x')$. The inner expectation is then the forced-next-target-action expectation in Equation (21). Since $c_t = \rho_t\lambda_t$, substitution into the preceding display gives

$$\begin{aligned} \mathbb{E}_\mu[R\widetilde{\rho}_t \mid x_t, a_t] &= (\rho_t - c_t)Q_t^\mu \\ &\quad + c_t\mathbb{E}\left[\sum_{a' \in \mathcal{A}_\mu(x_{t+1})} \pi(a' \mid x_{t+1})Q_{t+1}^c(x_{t+1}, a') \,\middle|\, x_t, a_t\right] \\ &= \rho_t Q_t^c. \end{aligned}$$

If $\rho_t = 0$, then $c_t = 0$ and both sides are zero. Conditioning each score summand on (x_t, a_t) and changing the current action from μ to π proves Equation (22); target-null actions have zero mass under that target-action expectation.

H.3 Proof of Proposition C.2 and Corollary C.3

Let $\mathcal{F}_t$ be the pre-action history specified in Proposition C.2. Exact change of measure for the current action gives

$$\mathbb{E}_\mu[\rho_t \mid \mathcal{F}_t] = \sum_{a \in \mathcal{A}_\mu(x_t)} \pi_\theta(a \mid x_t) = 1, \tag{47}$$

$$\mathbb{E}_\mu[\rho_t s_t \mid \mathcal{F}_t] = \sum_{a \in \mathcal{A}_\mu(x_t)} \nabla_\theta \pi_\theta(a \mid x_t) = 0. \tag{48}$$

The support and interior assumptions ensure that these ratios and scores are well defined.

We prove Equation (23) by backward induction. At $t = T-1$, $\widetilde{\rho}_{T-1} = \rho_{T-1}$, so the result follows from Equation (47). Suppose that $\mathbb{E}_\mu[\widetilde{\rho}_{t+1} - 1 \mid \mathcal{F}_{t+1}] = 0$. Both c_t and s_t are $\mathcal{F}_{t+1}$-measurable because they depend only on the current state–action pair. The trace recursion and the tower property therefore give

$$\begin{aligned}\mathbb{E}_\mu[\widetilde{\rho}_t - 1 \mid \mathcal{F}_t] &= \mathbb{E}_\mu[\rho_t - 1 \mid \mathcal{F}_t] \\ &\quad + \mathbb{E}_\mu[c_t \mathbb{E}_\mu[\widetilde{\rho}_{t+1} - 1 \mid \mathcal{F}_{t+1}] \mid \mathcal{F}_t] = 0.\end{aligned}$$

Using the same conditionally centered continuation term and Equation (48) yields

$$\begin{aligned}\mathbb{E}_\mu[\widetilde{\rho}_t s_t \mid \mathcal{F}_t] &= \mathbb{E}_\mu[\rho_t s_t \mid \mathcal{F}_t] \\ &\quad + \mathbb{E}_\mu[c_t s_t \mathbb{E}_\mu[\widetilde{\rho}_{t+1} - 1 \mid \mathcal{F}_{t+1}] \mid \mathcal{F}_t] = 0.\end{aligned}$$

This proves Proposition C.2. Notice that future ratios need not be independent of the current action: the required fact is the conditional mean-zero property after conditioning on the complete pre-action history at $t+1$.

For the corollary, applying the tower property to each token gives $\mathbb{E}_\mu[H_i \mid \mathcal{P}] = 0$. Conditional independence of the sampled responses implies that B_{-i} and H_i are independent given $\mathcal{P}$, so

$$\mathbb{E}_\mu[B_{-i} H_i \mid \mathcal{P}] = \mathbb{E}_\mu[B_{-i} \mid \mathcal{P}] \mathbb{E}_\mu[H_i \mid \mathcal{P}] = 0.$$

This proves Equation (25). Finally, the pathwise identity

$$R_i - \bar{R} = \frac{K-1}{K}(R_i - B_{-i})$$

proves Equation (26). The result concerns the prescribed score-attached semigradient H_i: the computed trace, reward, and baseline are held fixed, and parameter dependence is retained only through the current-action score attachment in Equation (36). It is not the derivative of the trace recurrence; differentiating its current or future ratios or its coefficients generally produces different terms.

I Time-augmented matrix representation

The finite-horizon operator identities also admit a compact matrix form. Time augmentation is important here: it makes the transition matrices strictly block upper triangular, avoiding an absorbing self-loop at discount factor one. This is a finite polynomial identity, not an infinite-horizon contraction argument.

At each time, restrict coordinates to $\mathcal{Z}_t^\mu$ and stack state-action functions into a vector. Let r be the terminal-payoff vector, with $r_t = 0$ for $t < T-1$ and $r_{T-1}(x,a) = Q_{T-1}^\mu(x,a)$. For $t < T-1$, define

$$(P^\nu F)_t(x,a) := \mathbb{E}\left[\sum_{a' \in \mathcal{A}_\mu(x_{t+1})} \nu(a' \mid x_{t+1}) F_{t+1}(x_{t+1}, a') \,\middle|\, x_t = x, a_t = a \right],$$

with a zero terminal block. Then $\mathcal{T}^\nu F = r + P^\nu F$. Write $D_f := \mathrm{diag}(f)$ for coordinatewise multiplication by a stacked function f, and set

$$P^{c\mu} := P^\mu D_c = P^\pi D_\lambda.$$

The second equality follows from $\mu c = \pi\lambda$ at the next action. Since $P^{c\mu}$ is nilpotent,

$$(I - P^{c\mu})^{-1} = \sum_{n=0}^{T-1} (P^{c\mu})^n \tag{49}$$

is a finite polynomial.

Proposition I.1 (Operator matrix identities) *For $V \in \mathbb{V}$, let $q = \mathcal{B}V$. Then*

$$\begin{aligned} \mathcal{R}_c^\pi q &= q + (I - P^{c\mu})^{-1}(\mathcal{T}^\pi q - q), \\ \mathcal{V}_c^{\mathrm{sa}} V &= D_\rho q + D_c (I - P^{c\mu})^{-1}(\mathcal{T}^\pi q - q) \\ &= D_\rho\big[(I - D_\lambda)q + D_\lambda[\mathcal{R}_c^\pi q]\big], \qquad \mathcal{B}(\mathcal{V}_c^{\mathrm{s}} V) = \mathcal{R}_c^\pi q. \end{aligned} \tag{50}$$

At the behavior value, these identities reduce to

$$Q^c = (I - D_\lambda)Q^\mu + D_\lambda[\mathcal{R}_c^\pi Q^\mu], \qquad \mathcal{V}_c^{\mathrm{sa}} V^\mu = D_\rho Q^c. \tag{51}$$

Proof. The Retrace recursion in Equation (29) stacks as

$$\mathcal{R}_c^\pi q - q = (\mathcal{T}^\pi q - q) + P^{c\mu}(\mathcal{R}_c^\pi q - q).$$

Inverting $I - P^{c\mu}$ with Equation (49) proves Equation (50). The remaining identities are the matrix forms of Theorem C.4 and Corollary G.3. □

Remark I.2 (Restriction on geometric simplifications) *For state-action-dependent traces, the inverse in Equation* (49) *is a polynomial in $P^{c\mu}$. It may be replaced by a polynomial in λP^π only for the constant-trace family $c = \lambda\rho$, where $P^{c\mu} = \lambda P^\pi$. The threshold family $c = \min\{\bar{c}, \rho\}$ does not generally admit that simplification.*

Theorem I.3 (State-action V-trace bounds) *Suppose the one-step target-policy backup improves the behavior action value coordinatewise:*

$$\mathcal{T}^\pi Q^\mu \geq Q^\mu. \tag{52}$$

Then, on the finite time-augmented behavior support,

$$\boxed{Q^\mu \leq Q^c \leq \mathcal{R}_c^\pi Q^\mu \leq Q^\pi.}$$

$$\boxed{\rho_t(x,a) Q_t^\mu(x,a) \leq [\mathcal{V}_c^{\mathrm{sa}} V^\mu]_t(x,a) \leq \rho_t(x,a) Q_t^\pi(x,a).} \tag{53}$$

The lower operator bound is attained at $c = 0$, and the upper operator bound at full suffix correction $c_j = \rho_j$ on the remaining suffix. Without Equation (52)*, no such numerical ordering is implied by the structural interpolation alone.*

Proof. Let $d := \mathcal{T}^\pi Q^\mu - Q^\mu \geq 0$. Equation (50) gives

$$\mathcal{R}_c^\pi Q^\mu - Q^\mu = \sum_{n=0}^{T-1} (P^{c\mu})^n d \geq 0.$$

Because $P^{c\mu} = P^\pi D_\lambda$ and $0 \leq \lambda \leq 1$, we have $0 \leq P^{c\mu} \leq P^\pi$ coordinatewise. Nonnegative-matrix induction therefore gives $(P^{c\mu})^n d \leq (P^\pi)^n d$. The finite-horizon Bellman identity $Q^\pi - Q^\mu = \sum_{n=0}^{T-1} (P^\pi)^n d$ proves $\mathcal{R}_c^\pi Q^\mu \leq Q^\pi$. Finally, Equation (51) expresses each Q_t^c as the coordinatewise convex combination $(1 - \lambda_t) Q_t^\mu + \lambda_t [\mathcal{R}_c^\pi Q^\mu]_t$, proving the middle inequalities. Multiplying $Q_t^\mu \leq Q_t^c \leq Q_t^\pi$ by $\rho_t(x,a) \geq 0$ and using Equation (45) proves Equation (53), including the $\rho_t = 0$ boundary case. □

J Further qualifications

The main development uses complete terminal-reward trajectories and treats the continuation target as fixed during the policy update. This section makes those two scope conditions precise, extends the construction to dense rewards, and records finite-horizon boundary checks that are useful in implementations.

J.1 Reward-only state-value V-trace and Retrace estimators

At the behavior-value inputs, separate complete-trajectory derivations give

$$\mathbb{E}_\mu[R\widetilde{\rho}_t \mid x_t] = [\mathcal{V}_c^{\mathrm{s}} V^\mu]_t(x_t),$$
$$\mathbb{E}_\mu[R\widetilde{\rho}_{t+1} \mid x_t, a_t] = [\mathcal{R}_c^\pi Q^\mu]_t(x_t, a_t), \qquad t < T-1,$$

with R itself estimating the terminal Retrace target. These are derived estimators, not substitutions of R for every value inside the residual definitions. They qualify any claim that earlier trace operators intrinsically require a learned critic. Canonical finite-unroll algorithms do learn and bootstrap critics; complete terminal trajectories permit the special Monte Carlo identities above.

J.2 Why a smaller cap does not help

Write h for the effective correction horizon, the expected number of tokens over which the trace still carries weight. For the constant-λ family $c_t = \lambda\rho_t$ it is $h = 1/(1-\lambda)$; for the threshold family $c_t = \min\{\bar{c}, \rho_t\}$ it is $1/(1-\mathbb{E}_\mu[\lambda_t])$ with $\lambda_t = \min\{1, \bar{c}/\rho_t\}$, so the cap $\bar{c}$ selects h.

Proposition J.1 (Variance is linear in the horizon) *Let the per-token log-ratios be i.i.d. with variance σ^2 and $\mathbb{E}_\mu[\rho_t] = 1$. For the constant-$\lambda$ family,*

$$\frac{\mathrm{Var}(\widetilde{\rho}_t)}{\mathrm{Var}(\rho_t)} = \frac{1}{1-\lambda^2} + O(\sigma^2) = \frac{h}{1+\lambda} + O(\sigma^2).$$

Proof. Expand $\widetilde{\rho}_t$ as in Equation (17) and drop the terminal atom, which is $O(\lambda^{T-t-1})$. The remaining weights $w_k = (1-\lambda)\lambda^k$ sum to one, and for i.i.d. log-ratios the partial products satisfy $\mathrm{Cov}(\rho_{t:t+j}, \rho_{t:t+k}) = (1+\mathrm{Var}(\rho_t))^{\min\{j,k\}+1} - 1$ exactly, since the two products share exactly $\min\{j,k\}+1$ factors and the remaining factors are independent with unit mean; dividing by $\mathrm{Var}(\rho_t) = O(\sigma^2)$ leaves $\min\{j,k\}+1+O(\sigma^2)$. Writing $\min\{j,k\}+1 = \sum_{m\geq 0} \mathbf{1}\{j \geq m\}\mathbf{1}\{k \geq m\}$ and summing,

$$\frac{\mathrm{Var}(\widetilde{\rho}_t)}{\mathrm{Var}(\rho_t)} = \sum_{j,k\geq 0} w_j w_k(\min\{j,k\}+1) + O(\sigma^2) = \sum_{m\geq 0}\Big((1-\lambda)\sum_{j\geq m}\lambda^j\Big)^2 + O(\sigma^2),$$

and the inner sum is λ^m, so the total is $\sum_{m\geq 0}\lambda^{2m} = 1/(1-\lambda^2)$. Substituting $h = 1/(1-\lambda)$ gives $h/(1+\lambda)$. The ratio is also checked numerically at $\lambda \in \{.5, .8, .9\}$ in the reproduction script. □

The bias side cannot be settled the same way, because it depends on where along the continuation Q_t^μ and Q_t^π diverge. We therefore state it as an assumption rather than a result.

Assumption J.2 (Uniform accrual) *The gap $G := Q_t^\pi - Q_t^\mu$ accrues evenly over the remaining $T-t$ tokens, so correcting h of them removes a fraction $h/(T-t)$ of it.*

The two must be compared through the mean squared error, so the bias enters squared.

Corollary J.3 (When no horizon helps) *Write $L = T-t$ for the remaining length, v_0 for the variance at $h = 0$ and $b > 0$ for the constant of Proposition J.1, whose residual dependence on h through $1/(1+\lambda)$ varies by at most a factor of two over $\lambda \in [0,1)$ and is absorbed into b. Under Assumption J.2,*

$$\mathrm{MSE}(h) = G^2\Big(1 - \frac{h}{L}\Big)^2 + v_0 + b\sigma^2 h,$$

whose derivative at $h = 0$ is $b\sigma^2 - 2G^2/L$. If that is nonnegative, MSE *is nondecreasing in h: no cap helps, and shrinking $\bar{c}$ changes only how fast the loss accrues. Otherwise* MSE *is minimized at the interior horizon $h^\star = L\big(1 - b\sigma^2 L/(2G^2)\big)$.*

So the question is decided at the margin, by whether $2G^2/L$ exceeds the per-token variance cost $b\sigma^2$ — a comparison that does not involve h, which is why no choice of $\bar{c}$ can change its sign. Our sweeps place it on the unfavorable side. Assumption J.2 is what makes the squared bias quadratic in h; if the gap were front-loaded the favorable regime would be reached sooner, and we did not test that (Section 7).

J.3 DENSE-REWARD EXTENSION

Suppose instead that action a_s is followed by an observed reward r_s. Replace the terminal-aware lift by the usual reward-aware backup $\mathbb{E}[r_s + V_{s+1}(x_{s+1}) \mid x_s, a_s]$ (with $V_T = 0$), and replace the nonterminal residual $\rho_s V_{s+1} - V_s$ by $\rho_s(r_s + V_{s+1}) - V_s$. The observed return tail is then a Monte Carlo sample of the corresponding behavior continuation, and the same finite telescoping argument still applies—but it does *not* reproduce a single multiplier on the return. Writing the partial trace

$$\widetilde{\rho}_{t:k} := \rho_t + \sum_{s=t+1}^{k} c_{t:s-1}(\rho_s - 1), \qquad \widetilde{\rho}_{t:T-1} = \widetilde{\rho}_t,$$

the pathwise realization at $V = V^\mu$ is

$$\sum_{k=t}^{T-1} r_k \, \widetilde{\rho}_{t:k},$$

each reward carrying only the trace accumulated up to its own arrival, rather than $\widetilde{\rho}_t \sum_{k \geq t} r_k$. The two coincide exactly when a single reward arrives at termination, which is the setting of the main text, and differ otherwise. An implementation therefore accumulates partial traces across reward-arrival steps and cannot apply one detached multiplier $w_t = \widetilde{\rho}_t / \rho_t$ to the return tail. Exact indexing still depends on whether rewards attach to actions or to transitions; the main text adopts a single terminal reward to keep that convention unambiguous.

J.4 THE FULL-SUFFIX ENDPOINT IS AN ANCHORED SEMIGRADIENT

Proposition J.4 (Reset-value objective has cross terms) *At $c = \rho$, the sample direction is*

$$\widehat{G}_{\text{suffix}} = R \sum_{t=0}^{T-1} \rho_{t:T-1} s_t.$$

It is not generally the full derivative of the self-consistent reset-value objective

$$\mathcal{J}_{\text{reset}}(\theta) := \sum_{k=0}^{T-1} \mathbb{E}_{x_k \sim d_k^\mu}[V_k^{\pi_\theta}(x_k)]. \tag{54}$$

Differentiation yields

$$\begin{aligned} \nabla_\theta \mathcal{J}_{\text{reset}}(\theta) &= \sum_{k=0}^{T-1} \sum_{t=k}^{T-1} \mathbb{E}_\mu[R \rho_{k:T-1} s_t] \\ &= \mathbb{E}_\mu[\widehat{G}_{\text{suffix}}] + \sum_{0 \leq k < t \leq T-1} \mathbb{E}_\mu[R \rho_{k:T-1} s_t]. \end{aligned} \tag{55}$$

The cross terms need not vanish, so the two directions are unequal in general.

Proof. For each reset time k, the likelihood-ratio identity differentiates every target-policy action from k through $T-1$:

$$\nabla_\theta V_k^{\pi_\theta}(x_k) = \mathbb{E}_{a_{k:T-1} \sim \pi_\theta} \left[R \sum_{t=k}^{T-1} s_t \,\middle|\, x_k \right].$$

Changing the suffix law from μ to π yields Equation (55). The diagonal terms $k = t$ are exactly $\mathbb{E}_\mu[\widehat{G}_{\text{suffix}}]$; the pairs $k < t$ are the displayed cross terms. This is why the main text uses the frozen-target surrogate in Equation (32), rather than treating full suffix correction as the gradient of Equation (54). □

Remark J.5 (A two-step nonzero cross term) *Let $T = 2$, let the first-action policy be independent of θ, and let the second action be Bernoulli with $\pi_\theta(a_1 = 1 \mid x_1) = \sigma(\theta)$ at every history. Set $R = a_1$ and take any full-support behavior policy. At $\theta = 0$,*

$$\mathbb{E}_\pi[R s_1] = \sigma(0)(1 - \sigma(0)) = \tfrac{1}{4}.$$

The full-suffix direction contains this derivative once. The reset-value objective contains it once for the reset at $k = 0$ and once for the reset at $k = 1$, and therefore has derivative $1/2$. Thus the missing cross term is nonzero even in the smallest nontrivial horizon.

J.5 Boundary checks

With one remaining action, $\widetilde{\rho}_t = \rho_t$ and $Q^c_t = Q^\mu_t = Q^\pi_t$. With two remaining actions and constant λ,

$$\widetilde{\rho}_t = (1-\lambda)\rho_t + \lambda\rho_{t:t+1}.$$

The behavior- and target-continuation weights are therefore $(1-\lambda, \lambda)$, not $(1-\lambda^2, \lambda^2)$; the coin is tossed after the current action and exactly once before the next action. With three remaining actions,

$$\widetilde{\rho}_t = (1-\lambda)\rho_t + (1-\lambda)\lambda\rho_{t:t+1} + \lambda^2\rho_{t:t+2}.$$

More generally, the coefficients in Equation (17) sum to

$$(1-\lambda)\sum_{k=0}^{T-t-2}\lambda^k + \lambda^{T-t-1} = 1.$$

The final term is the finite-horizon no-switch atom. Omitting it produces a sub-probability mixture; squaring the two-step switch probability corresponds to mixing an already mixed switching value a second time. These checks expose both errors directly.

K Conceptual illustration of the finite-horizon mixture

Figure 3 visualizes the constant-λ expansion. It complements the switch-timing diagram in the main text without introducing an additional assumption.

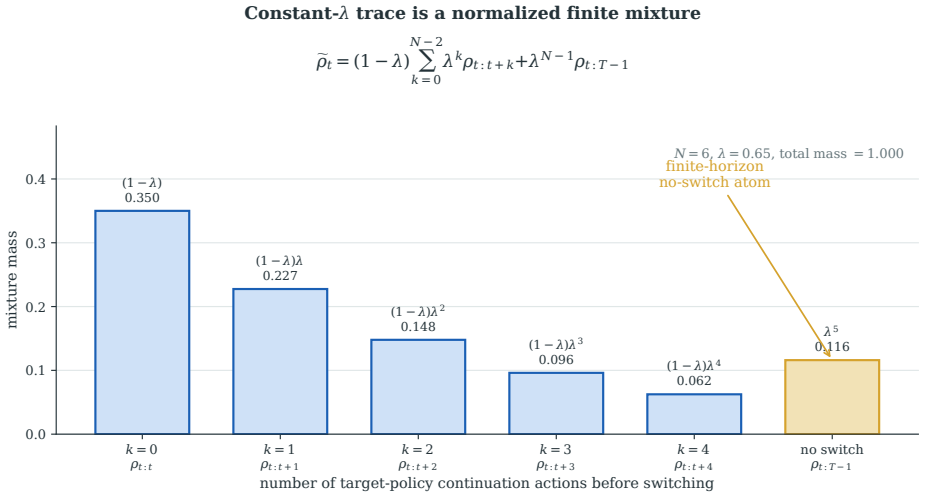

Figure 3: Finite-horizon constant-λ mixture with six remaining actions and $\lambda = 0.65$. Each blue bar is the probability of switching after a finite number of target-policy continuation actions. The gold bar is the no-switch atom λ^5; it is required for the mixture weights to sum to one.

L Experiments

L.1 Questions and protocol

The decomposition motivates five empirical tests. First, when $\rho_t = 1$, all trace endpoints coincide algebraically; an age-zero control tests how closely the implementation approaches that regime. Second, under rollout reuse, a transformed-trace factorial tests whether the response to the PPO clipping radius depends on the trace-coefficient cap. Third, as that cap approaches full-suffix correction, we measure the predicted growth in importance-weight tails. Fourth, complete-recipe learning curves test whether the original V-Trace PO-TC recipe's advantage over standard GRPO changes with reuse. Fifth, the matched-clipping study holds $\epsilon = .2$ fixed to isolate suffix correction

under local clipping and local versus transformed clipping at $\bar{c} = 1$. These questions remain separate rather than being collapsed into one endpoint score.

All experiments use a pinned Qwen3-8B-Base checkpoint (Yang et al., 2025) and MATH training data (Hendrycks et al., 2021). Each fresh batch contains 32 prompts and eight independently sampled responses per prompt. Rollouts are generated synchronously by vLLM (Kwon et al., 2023) at temperature one; the stored sampler log-probability, not a recomputed actor probability, is the behavior denominator. Training responses are capped at 2,048 tokens. Both reward-side KL shaping and actor-loss KL regularization are disabled, and all arms use norm-one global gradient clipping.

We report two frozen core evaluation suites separately. The primary recent-competition endpoint contains all 150 problems from AIME 2025, AIME 2026, BRUMO 2025, HMMT February 2025, and HMMT November 2025. Canonical MATH-500 is a separate 500-problem key benchmark endpoint and is never pooled into the headline recent-competition score. We additionally report all 30 AIME 2024 problems as a standalone, familiar secondary benchmark, using the `aime24` slice of the compiled test-set release accompanying Shafayat et al. (2025); that work attributes the source benchmark to Jia (2024). It is not included in either core suite or in their pooled score, and we make no post-cutoff or contamination-free claim for it. The common protocol uses eight responses per problem, seed 2026, temperature one, and a 4,096-token response limit; strict sampled pass@1 is the mean binary verdict from `math-verify==0.8.0`. A pooled 650-problem score over only the recent suite and MATH-500 is descriptive because its weighting is dominated by MATH-500. The factorial, pilot and matched-update experiments comprise 92 frozen model targets evaluated under this common protocol. The earlier complete-recipe learning curves add 240 checkpoint targets, and the matched-clipping study adds 360; every target is evaluated on both core endpoints and, separately, on AIME 2024. AIME 2024 enters no pooled score or primary test. An operator-preserving normalization audit finds no exact prompt overlap between either core suite and the MATH training split, but does not rule out semantic duplication or pretraining exposure.

This evaluation replaces an earlier benchmark pool for data-eligibility reasons. Its composition, checkpoint set, sampling protocol, and analyses were frozen before replacement scores were generated, but the earlier endpoint's outcomes had already been inspected; we therefore do not present the replacement as an untouched confirmation. No score from the superseded pool enters the analysis. The training seed remains the experimental unit, and all headline comparisons use fixed final checkpoints rather than best-checkpoint selection. Appendix M gives the complete data provenance, optimizer and advantage construction, truncation convention, overlap audit, and statistical protocol.

L.2 Trace correction changes the empirical effect of PPO clipping

Confirmatory design. We test the interaction between suffix correction and PPO clipping in the high-reuse regime $E = 8$, where each fresh rollout batch supports eight consecutive actor updates. Reuse age zero is the first update on a fresh batch; age k follows k previous updates on that batch, so age seven is the final update at $E = 8$. Before inspecting these outcomes, we fixed a 3×2 factorial crossing $\bar{c} \in \{0, 1, 10^7\}$ with $\epsilon \in \{0.2, 1\}$ and six training seeds (2027–2032). Every cell uses the original transformed-trace clipping placement. The value 10^7 is a large-cap proxy for full-suffix importance sampling, not literal infinity. Every run starts from the same Qwen3-8B-Base checkpoint and trains for 50 fresh batches, or 400 actor updates, using 12,800 sampled responses. At the fixed final checkpoint, the primary accuracy endpoint is sampled pass@1 on the 150-problem recent suite; canonical MATH-500 sampled pass@1 is reported separately as a key benchmark endpoint. Both use eight responses per problem under the common evaluation protocol above. AIME 2024 is a standalone descriptive secondary endpoint. The training seed is the experimental unit; the exploratory seed used to formulate the hypothesis is excluded.

All 36 training runs and their primary endpoints completed and passed the completeness audit. The prespecified interaction for seed s is

$$D_s := \left[Y_{s,\bar{c}=1,\epsilon=1} - Y_{s,\bar{c}=1,\epsilon=.2}\right] - \left[Y_{s,\bar{c}=0,\epsilon=1} - Y_{s,\bar{c}=0,\epsilon=.2}\right].$$

It is positive in every seed, with mean $+5.36$ percentage points, Student-t 95% CI $[0.98, 9.74]$, and exact two-sided sign-flip $p = 0.03125$ (Figure 4a). Descriptively, the mean interaction is positive in each of the five constituent benchmark sources. The interaction is easy to interpret from the cell means in Table 5: widening ϵ changes recent-competition pass@1 by -3.42 points at $\bar{c} = 0$ but by $+1.94$ points at $\bar{c} = 1$. The latter within-cap contrast has 95% CI $[0.76, 3.13]$, although it does not

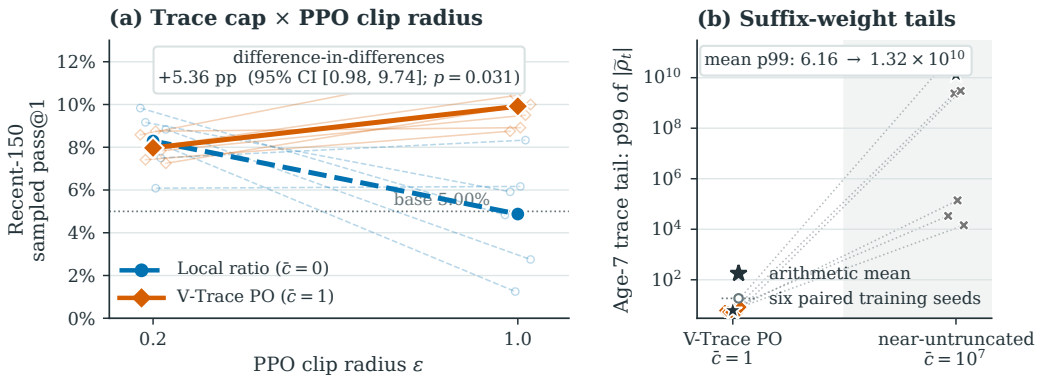

Figure 4: **Unit-capped continuation coefficients reverse the empirical effect of widening the PPO interval under eight-fold reuse.** (a) Held-out strict sampled pass@1 at the fixed final checkpoint. Thin lines pair the two clipping radii within each training seed and thick lines show six-seed means. Widening ϵ hurts the local-ratio endpoint $\bar{c} = 0$ but helps V-TRACE PO-TC with $\bar{c} = 1$, yielding the prespecified $+5.36$ percentage-point difference-in-differences on the 150-problem recent-competition suite. The horizontal line is the 5.00% frozen base-model endpoint. (b) The corresponding reuse-age-seven suffix-weight p99 is well controlled by $\bar{c} = 1$ but spans many orders of magnitude for the large-cap proxy. Small markers are seed means over training batches and stars are means across seeds.

Table 5: **Six-seed recent-competition accuracy and mechanism summary at $E = 8$.** Sampled pass@1 uses eight responses per problem and is the fixed-checkpoint mean $\pm$ sample standard deviation (percent). Mechanisms are measured at reuse age seven, averaged first over the 50 training batches within each run and then across seeds. Coefficient ESS is normalized to $[0, 1]$; $|\delta^{\text{fp}}|$ is the absolute log ratio after the implementation's $[-20, 20]$ safeguard; L1 suppressed is the fraction of absolute detached surrogate-coefficient mass removed by PPO clipping, not full parameter-gradient bias. The frozen base-model pass@1 is 5.00%.

$\bar{c}$	ϵ	Pass@1	Mean $\lvert\delta^{\text{fp}}\rvert$	Coeff. ESS	L1 suppressed	Trace p99	Grad. norm
0	.2	8.29 ± 1.34	.0374	.488	.061	1.23	.619
0	1	4.88 ± 2.54	.5925	.252	.066	3.53	4.56
1	.2	7.97 ± 0.60	.0105	.386	.621	2.49	.0558
1	1	9.92 ± 1.17	.0136	.248	.351	6.16	.0771
10^7	.2	8.64 ± 0.31	.0088	.258	.580	1.24×10^3	.0408
10^7	1	9.32 ± 0.87	.0114	.150	.388	1.32×10^{10}	.0588

reject after the prespecified Holm correction ($p_{\text{Holm}} = 0.125$). All secondary confidence intervals are unadjusted Student-t intervals; only the exact p-values receive Holm adjustment.

For scale, the frozen base model scores 5.00%, whereas the $\bar{c} = 1, \epsilon = 1$ condition averages 9.92% after training. This is a descriptive before–after comparison, not the factorial estimand. The low held-out baseline does not imply sparse training feedback at initialization: across the six distinct pre-update MATH rollout batches (256 responses each), the binary verifier success rate averages 45.96% (range 35.94–55.86%). We did not log this rate over the rest of training. The pass-any and total-success counts further imply that at least 88 of the 192 prompt groups (45.83%; 13–16 of 32 in every batch) contain both correct and incorrect samples, and therefore have nonzero within-group GRPO advantages.

The primary result is an interaction, not a post-hoc declaration that one cell wins universally. The practical comparison between $\bar{c} = 1, \epsilon = 1$ and the matched current-token PPO/GRPO-style baseline ($\bar{c} = 0, \epsilon = 0.2$) is $+1.63$ points, with 95% CI $[0.44, 2.81]$. Its exact paired sign-flip test is nevertheless $p = 0.0625$ ($p_{\text{Holm}} = 0.1875$), so the descriptive Student-t interval should not be read as an exact-test rejection. Likewise, the direct comparison between $\bar{c} = 1, \epsilon = 1$ and the large-cap $\epsilon = 1$ cell is $+0.60$ points, with 95% CI $[-0.94, 2.13]$ and $p_{\text{Holm}} = 0.375$. The accuracy data

Table 6: Canonical MATH-500 sampled pass@1 with eight responses per problem for the same six-seed 3×2 factorial (mean $\pm$ sample standard deviation, percent). MATH-500 is reported separately from the recent-competition primary endpoint; the frozen base model scores 43.93%.

$\bar{c}$	$\epsilon = .2$	$\epsilon = 1$
0	61.35 ± 3.45	44.04 ± 17.02
1	67.76 ± 2.40	73.50 ± 3.71
10^7	71.03 ± 3.55	72.76 ± 2.42

Table 7: Standalone AIME 2024 results for the six-seed $E = 8$ factorial. Sampled pass@1 and observed pass@8 are mean $\pm$ sample standard deviation across training seeds (percent); cap is the mean 4,096-token cap rate. Pass@8 and every comparison on this 30-problem endpoint are descriptive.

$\bar{c}$	ϵ	Pass@1	Pass@8	Cap
0	.2	12.71 ± 4.30	28.33 ± 9.37	27.08
0	1	2.78 ± 4.00	9.44 ± 11.04	65.97
1	.2	14.79 ± 0.68	35.00 ± 7.53	8.61
1	1	16.18 ± 1.98	40.56 ± 5.34	9.44
10^7	.2	14.10 ± 2.81	30.00 ± 6.67	5.14
10^7	1	15.35 ± 2.13	34.44 ± 7.20	6.39
Base	–	5.42	26.67	4.58

therefore establish that unit-capped continuation coefficients respond more favorably to a wider clipping radius in this regime; they do not establish a total ordering of trace-coefficient caps.

At the common wide interval, exploratory paired contrasts show consistent separation from the local-ratio endpoint. Every cell has the same generation budget—50 rollout batches and 12,800 responses—and 400 actor updates. Direct decompositions of the factorial give $(\bar{c}{=}1, \epsilon{=}1) - (\bar{c}{=}0, \epsilon{=}1) = +5.04$ points (95% CI $[+1.68, +8.40]$) and $(\bar{c}{=}10^7, \epsilon{=}1) - (\bar{c}{=}0, \epsilon{=}1) = +4.44$ points ($[+1.34, +7.55]$); both exact paired sign-flip tests have unadjusted $p = 0.03125$, with all six paired differences positive in both comparisons. Thus the data support a scoped matched-budget advantage for both multi-step corrections over current-token correction in this high-reuse, wide-update setting. These direct decompositions were not members of the prespecified Holm family and are not a confirmatory ordering across clipping radii: at $\epsilon = .2$, the recent-competition means are 8.29%, 7.97%, and 8.64% for $\bar{c} \in \{0, 1, 10^7\}$.

The separately reported canonical MATH-500 benchmark shows the same wide local-cell failure (Table 6). Its trace-threshold-by-clipping-radius interaction is $+23.05$ points (95% CI $[7.94, 38.15]$, exact $p = 0.03125$). The tuned $(\bar{c} = 1, \epsilon = 1)$ arm exceeds the standard $(\bar{c} = 0, \epsilon = .2)$ arm by 12.16 points ($[6.16, 18.15]$, exact $p = 0.03125$). At $\epsilon = 1$, the unit-capped and large-cap arms exceed $\bar{c} = 0$ by 29.46 points ($[10.41, 48.52]$) and 28.72 points ($[9.99, 47.46]$), respectively; both exact tests have $p = 0.03125$. As with their Recent-150 counterparts, these decompositions are not members of the prespecified Holm family and are exploratory; four tests on the same six seeds would not survive a Holm correction over that set. The unit-capped arm is again best numerically, but its $+0.74$-point difference from the wide large-cap arm is inconclusive ($[-5.01, 6.50]$, exact $p = 0.75$).

Standalone AIME 2024. The familiar 30-problem secondary endpoint gives a compatible descriptive picture (Table 7). The frozen base scores 5.42% sampled pass@1. Among the $E = 8$ cells, $(\bar{c} = 1, \epsilon = 1)$ has the highest mean pass@1 (16.18%) and pass@8 (40.56%). Relative to the standard $(\bar{c} = 0, \epsilon = .2)$ cell, its paired descriptive differences are $+3.47$ and $+12.22$ points, respectively, with positive differences in five of six seeds; relative to the wide large-cap cell they are $+0.83$ and $+6.11$ points. These are descriptive comparisons on only 30 shared problems, not additional hypothesis tests. In particular, the collapse of the wide $\bar{c} = 0$ cell coincides with a 65.97% response-cap rate and should not be interpreted as a correction-only effect.

The three-seed $E = 1$ control is less separated: its four pass@1 means range from 10.42% to 11.81%. Along the six-seed long-horizon $E = 1$ GRPO curve ($\bar{c} = 0$, $\epsilon = .2$; 100, 200 and 400 fresh batches,

so up to 102,400 generated responses), pass@1 is 11.32%, 14.51% and 18.40%, versus 16.18% for the fixed 50-batch, 400-update $E = 8$ V-TRACE PO-TC recipe ($\bar{c} = 1$, $\epsilon = 1$), which generates 12,800 responses. "Matched-update" therefore means eight times the rollout budget for the $E = 1$ arm. Thus AIME 2024 supports an early descriptive rollout-efficiency signal but not eight-fold matched-update dominance. Table 7 reports every factorial cell; pass@8 and scorer diagnostics remain descriptive.

Why does the useful clipping range change? The training diagnostics support a coherent, though not causal, mechanism. At $\bar{c} = 1$, widening ϵ reduces the absolute coefficient mass suppressed by PPO clipping from 0.621 to 0.351, while mean $|\delta_t^{\mathrm{fp}}|$ changes only from 0.0105 to 0.0136 and the mean pre-global-clipping gradient norm remains small (0.0558 to 0.0771). In these runs, the wider interval therefore retains more of the suffix-corrected coefficient while measured mismatch and gradient norms remain small. At $\bar{c} = 0$, the same change increases mean $|\delta_t^{\mathrm{fp}}|$ from 0.0374 to 0.5925 and the gradient norm from 0.619 to 4.56, alongside the decline in recent-competition pass@1. These correlations do not by themselves identify a causal decomposition, but they match the hypothesis that the local ratio becomes fragile when a wider update range is combined with repeated reuse.

Truncation and PPO clipping solve different problems. At $\epsilon = 1$, $\bar{c} = 1$ is $+0.60$ recent-competition points above the large-cap proxy, with 95% CI $[-0.94, 2.13]$; this interval is inconclusive, not evidence of equivalence. Yet the large-cap reuse-age-seven trace p99 is 1.32×10^{10}, compared with 6.16 at $\bar{c} = 1$, and its coefficient ESS is 0.150 rather than 0.248 (Figure 4b). Its small logged aggregate gradient norm is therefore consistent with the outer PPO mask shielding optimization from an extreme raw tail, not with a well-conditioned estimator. In the three-seed unclipped stress test, all $\bar{c} = 1$ runs remain finite, although their mean age-seven gradient norm is already 1.95×10^3; all three large-cap runs produce at least one nonfinite scoped diagnostic. Among finite values logged before failure, the corresponding means are 1.46×10^{17} for the gradient norm and 5.54×10^{29} for trace p99. Capping each continuation coefficient at one sharply moderates, but does not bound, the cumulative suffix-weight tail; the outer PPO clip supplies an additional safety layer that is essential in these three stress runs. This orders-of-magnitude separation is consistent with the additive pathwise control in Proposition B.3.

Clipping-placement pilot. A separate single-seed $E = 8$ study compares transformed-trace PPO, local-ratio PPO with the detached continuation multiplier outside the clip, and the unclipped V-TRACE PG objective. The objectives agree when their PPO masks are inactive, but diverge sharply once suffix weights spread: the local-clipping large-cap run reaches a maximum logged pre-clipping gradient norm of 1.93×10^{14}, whereas the unclipped large-cap run stops after 41 of 50 batches on nonfinite traces. These results establish a stability distinction but do not rank the two clipped placements at $\bar{c} = 1$: their endpoint ordering changes across the two benchmarks and only one training seed was used. They motivated the replicated matched-clipping comparison below; V-TRACE PG remains a mechanism stress probe. Appendix M reports the complete pilot matrix, fixed checkpoint endpoints, and numerical safeguards.

L.3 REPLICATED MATCHED-CLIPPING STUDY

This study is superseded in the main text by the radius-by-placement factorial of Section 6.1, which repeats the same three arms and reuse depths at a second clipping radius and under mean-only advantages. The two agree on the quantity that matters here: the per-doubling slope of the gap over GRPO is $+0.298$ points in this study and $+0.328$ in the factorial's $\epsilon = .2$ half, measured on disjoint seeds under different advantage normalization (Appendix M.14). We keep the study on record because it is the independent replication behind that agreement, and because its per-doubling slope is the one contrast in either experiment that reaches $p = .03125$.

The pilot motivated two questions: whether unit-capped suffix correction helps when the ordinary local PPO mask is held fixed, and whether local or transformed-trace clipping is preferable at the same cap and radius. Before launch, we froze a $4 \times 3 \times 6$ design with $E \in \{1, 2, 4, 8\}$, seeds 2051–2056, and three $\epsilon = .2$ arms:

Arm	$\bar{c}$	PPO-clipped quantity	Role
Local GRPO	0	ρ_t	local-ratio control
V-TRACE PO	1	ρ_t	suffix-correction effect
V-TRACE PO-TC	1	$\hat{\rho}_t$	clipping-placement comparison

Each of the 72 runs starts from the same Qwen3-8B-Base checkpoint and uses 50 fresh MATH batches, 32 prompts per batch, eight responses per prompt, and $50E$ actor-update slots. The sampled batch and behavior log-probabilities remain fixed through its E epochs; the learner ratio and trace are recomputed after every update. Checkpoints are fixed at fresh-batch indices 10, 20, 30, 40, and 50. The shared base-model evaluations supply one unreplicated step-zero coordinate.

Writing $Y_{a,E,s}$ for Recent-150 normalized curve AUC, the primary suffix-correction interaction is

$$[Y_{\mathrm{Local},8,s} - Y_{\mathrm{GRPO},8,s}] - [Y_{\mathrm{Local},1,s} - Y_{\mathrm{GRPO},1,s}],$$

and the primary clipping-placement interaction replaces GRPO by Trace in the same expression, where Trace denotes V-TRACE PO-TC. Their means are $+0.981$ percentage points (95% Student-t CI $[0.226, 1.735]$, exact sign-flip $p = .0625$) and $+0.803$ points ($[0.010, 1.596]$, $p = .09375$), respectively. Both estimates favor Local more strongly at high reuse, but neither preregistered exact test rejects at 5%.

At $E = 8$, the secondary Local-minus-GRPO and Local-minus-Trace AUC gaps are $+1.114$ points ($[0.383, 1.845]$) and $+1.376$ points ($[0.614, 2.138]$), each with exact $p = .03125$ and all six seed pairs positive. Across all four reuse levels, the secondary Local-minus-GRPO AUC-gap slope is $+0.298$ points per doubling of E ($[0.067, 0.530]$, exact $p = .03125$). Figure 5 shows every checkpoint.

Table 8: **Aggregate matched-clipping learning curves.** Entries are mean normalized AUC / step-50 pass@1 in percentage points over six seeds. Bold marks the largest value within each benchmark and metric.

Benchmark	E	Local GRPO	V-TRACE PO	V-TRACE PO-TC
Recent-150	1	6.283/7.611	6.417/7.528	5.843/7.208
	2	7.461/8.972	7.469/9.167	7.057/8.236
	4	8.247/9.583	8.299/**10.431**	7.449/8.347
	8	7.189/7.500	**8.303**/9.111	6.926/7.542
MATH-500	1	59.698/70.367	59.900/70.717	57.212/68.275
	2	66.496/75.346	65.963/74.942	63.770/72.308
	4	**70.133**/77.592	69.530/**77.829**	64.969/73.033
	8	60.817/58.271	65.665/70.242	61.647/67.304
AIME 2024 (standalone)	1	9.396/11.319	9.882/12.986	8.479/10.764
	2	11.146/13.264	10.708/11.667	10.285/12.569
	4	12.090/14.792	12.306/15.278	11.854/12.431
	8	12.215/11.597	**13.924**/**15.347**	11.493/12.847

The primary Recent-150 AUC is largest for V-TRACE PO at $E = 8$, whereas its largest final endpoint occurs at $E = 4$; the result is robustness to aggressive reuse, not monotonic improvement with E. MATH-500 is secondary: GRPO at $E = 4$ has the largest AUC, while V-TRACE PO at $E = 4$ has the largest endpoint. AIME 2024 is descriptive and never pooled. Within-E curves and endpoints, log-linear reuse trends, and MATH-500 analogues do not replace the two primary Recent-150 interactions.

The fail-closed audit accepts all 72 training runs, all 360 retained checkpoints, and 720 logical held-out evaluations. Of these, 712 use their original outputs and eight Approved-650 targets use protocol-identical infrastructure repairs; repairs are source substitutions for the same frozen targets, not additional observations. The design is matched on fresh batches and sampled completions, not on optimizer updates, FLOPs, compute, or wall time, and it does not randomize realized policy mismatch. No large-cap local-clipping arm is trained because its detached future-ratio product would remain outside the PPO clip.

L.4 NEAR-ON-POLICY MECHANISM CONTROL

The same original-transformed-trace 2×2 comparison of $\bar{c} \in \{0, 1\}$ and $\epsilon \in \{0.2, 1\}$ was repeated for three seeds 2027–2029 with $E = 1$. Each arm uses 50 fresh batches, 50 actor updates, and

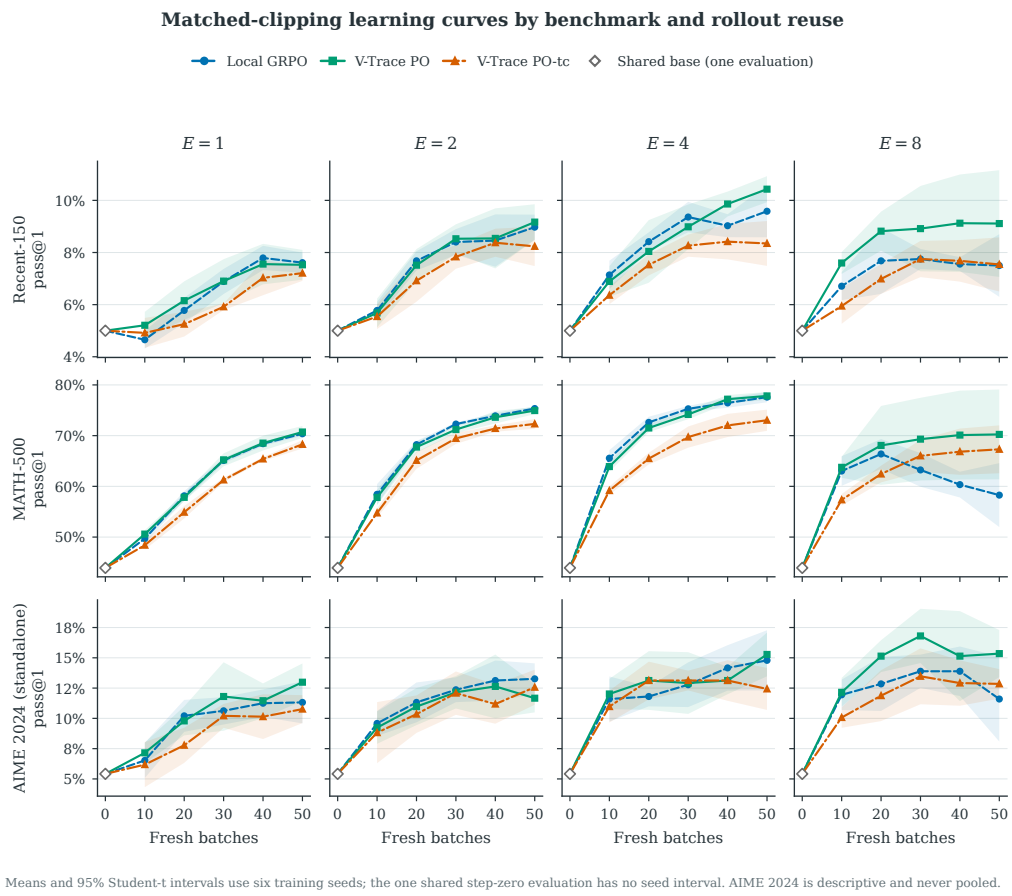

Figure 5: **Matched-clipping learning curves by benchmark and rollout reuse.** Every arm uses $\epsilon = .2$. Lines show mean sampled pass@1 and bands show 95% Student-t intervals across six training seeds at fresh batches 10–50. The batch-zero diamond is one shared frozen-base evaluation, not six seed replicates. Recent-150 is primary, MATH-500 is secondary, and AIME 2024 is standalone and descriptive.

12,800 sampled responses. Each fresh rollout batch then supports a single actor update, so reuse-induced policy lag is absent; small residual ratio deviations can still arise from sampler–trainer numerical differences. In this control, widening ϵ changes the $\bar{c} = 0$ mean $|\delta_t^{\text{fp}}|$ by only 0.0008 and the response-cap fraction by 0.0013, compared with changes of 0.555 and 0.344 at $E = 8$. For $(\bar{c}, \epsilon) \in \{(0, .2), (0, 1), (1, .2), (1, 1)\}$, the four recent-competition pass@1 means are respectively 8.28 ± 0.97, 7.47 ± 0.27, 7.25 ± 0.55, and 7.86 ± 0.63. The resulting three-seed interaction is $+1.42$ points (Student-t 95% CI $[0.06, 2.77]$), but with only three paired signs its exact two-sided p-value is 0.25. On MATH-500 the same four cell means are 71.50 ± 0.82, 68.34 ± 0.94, 68.12 ± 1.05, and 69.61 ± 0.77, and the interaction is $+4.65$ points ($[-2.54, 11.84]$, exact $p = 0.25$). Thus the large mismatch response in Table 5 appears in the observed high-reuse runs and is strongly attenuated at $E = 1$, rather than being an automatic consequence of changing ϵ. Because this control has three seeds and was designated secondary, we do not treat its accuracy contrast as an independent confirmation or as a formal cross-E interaction; the comparison does not separate reuse from cumulative optimizer exposure.

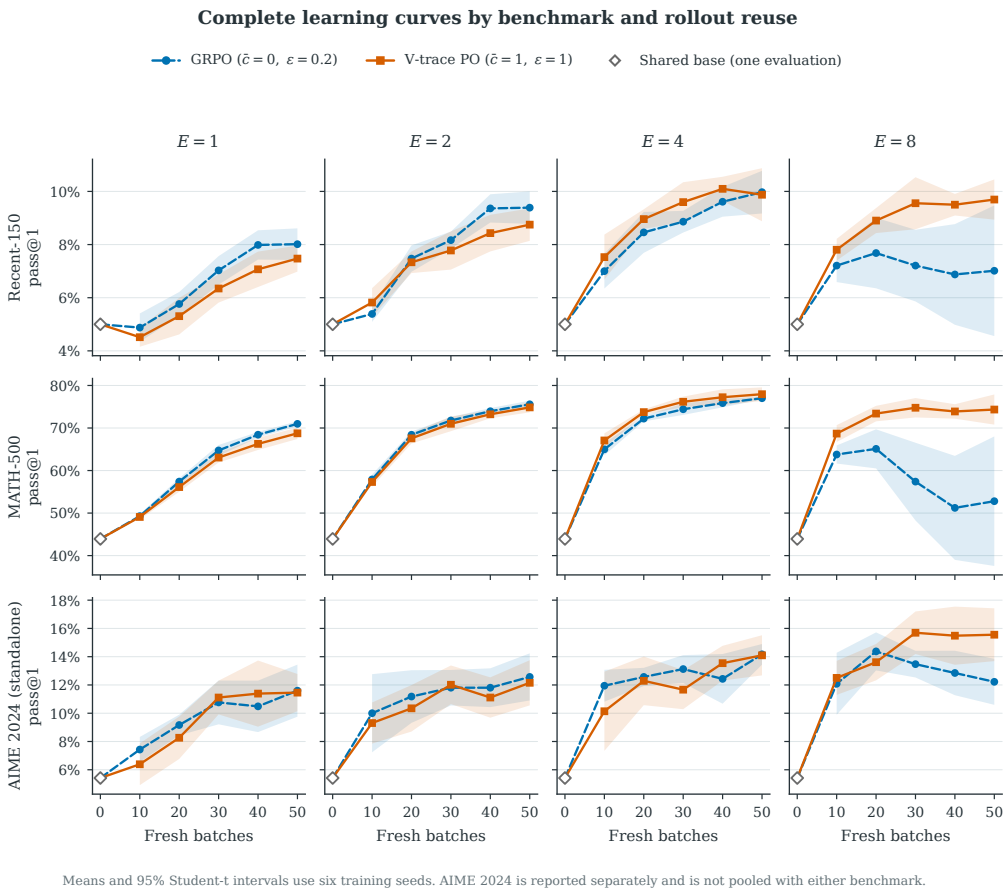

Figure 6: **Complete six-seed learning curves across rollout reuse.** Lines show mean sampled pass@1 and bands show two-sided Student-t 95% intervals across training seeds at batches 10–50. Diamonds at batch zero are the single shared frozen-base evaluation (5.00% on Recent-150, 43.93% on MATH-500, and 5.42% on AIME 2024), not six replicated base evaluations; no uncertainty band is attached to them. Every column uses the same 50 fresh-batch budget, while its optimizer-update budget is $50E$. AIME 2024 is a standalone descriptive endpoint and is never pooled with the two core benchmarks.

L.5 Complementary complete-recipe curves across rollout reuse

As complementary triangulation, we retain the earlier comparison of two complete, practically motivated recipes across $E \in \{1, 2, 4, 8\}$: standard GRPO-style training ($\bar{c} = 0, \epsilon = .2$) and the original V-Trace PO-tc ($\bar{c} = 1, \epsilon = 1$). For each reuse factor and recipe, six new training seeds (2033–2038) start from the same frozen base checkpoint and process 50 fresh batches. Checkpoints after batches 10, 20, 30, 40, and 50 are evaluated on Recent-150, MATH-500, and standalone AIME 2024. Thus the comparison is paired within seed and holds generated training data count fixed at 12,800 responses per run, but increasing E increases actor updates from 50 to 400. The two recipes differ in both $\bar{c}$ and ϵ; this study therefore estimates a recipe contrast, not a trace-only effect.

Figure 6 shows the complete curves. For a scalar summary that uses every checkpoint, let normalized curve AUC be the trapezoidal average of pass@1 over the fresh-batch coordinate from zero to 50. The preregistered Recent-150 estimand is the within-seed change in recipe gap from $E = 1$ to $E = 8$,

$$\left[\mathrm{AUC}_{\mathrm{VTP},8} - \mathrm{AUC}_{\mathrm{GRPO},8}\right] - \left[\mathrm{AUC}_{\mathrm{VTP},1} - \mathrm{AUC}_{\mathrm{GRPO},1}\right].$$

Here VTP abbreviates the original V-Trace PO-tc recipe ($\bar{c} = 1$, $\epsilon = 1$). Its mean is $+2.16$ pass@1 percentage points, with Student-t 95% CI $[+0.86, +3.47]$. Five of six seed-level interactions are positive, giving an exact two-sided sign-flip $p = 0.0625$. The positive t interval and the exact test should be reported together: the result is evidence that the relative curve advantage grows with reuse,

but it does not cross the conventional 0.05 threshold under the exact test. The remaining curve-study intervals and exact tests are secondary and unadjusted; they do not define additional familywise confirmatory rejections.

Table 9: **Aggregate replicated learning-curve results.** AUC is normalized trapezoidal pass@1 over fresh batches 0–50 and is expressed in percentage points. The last two columns are paired V-TRACE PO-TC-minus-GRPO differences with two-sided Student-t 95% intervals over six training seeds; "step 50" is the fixed final checkpoint. AIME 2024 rows are descriptive.

Benchmark	E	GRPO AUC	V-TRACE PO-TC AUC	Paired ΔAUC [95% CI]	Paired Δ step 50 [95% CI]
Recent-150	1	6.43	5.89	-0.54 $[-0.78, -0.30]$	-0.54 $[-1.20, +0.12]$
	2	7.52	7.25	-0.27 $[-0.57, +0.03]$	-0.64 $[-1.53, +0.25]$
	4	8.28	8.72	$+0.44$ $[+0.25, +0.63]$	-0.10 $[-1.35, +1.15]$
	8	7.00	8.62	$+1.63$ $[+0.45, +2.80]$	$+2.68$ $[-0.24, +5.60]$
MATH-500	1	59.46	58.17	-1.28 $[-2.21, -0.36]$	-2.22 $[-3.52, -0.93]$
	2	66.36	65.70	-0.66 $[-1.77, +0.45]$	-0.73 $[-1.61, +0.16]$
	4	69.58	71.03	$+1.45$ $[+0.41, +2.48]$	$+0.95$ $[-0.74, +2.63]$
	8	57.17	69.97	$+12.81$ $[+5.73, +19.89]$	$+21.55$ $[+4.76, +38.34]$
AIME 2024	1	9.27	9.12	-0.15 $[-0.65, +0.35]$	-0.14 $[-3.00, +2.72]$
(standalone)	2	10.76	10.31	-0.44 $[-1.08, +0.19]$	-0.42 $[-2.52, +1.69]$
	4	11.97	11.48	-0.49 $[-1.63, +0.64]$	-0.07 $[-1.12, +0.98]$
	8	12.32	13.56	$+1.24$ $[+0.46, +2.02]$	$+3.33$ $[+1.00, +5.66]$

At $E = 8$, the Recent-150 AUC gap is $+1.63$ points ($[+0.45, +2.80]$, exact sign-flip $p = 0.0625$), while the batch-50 gap is $+2.68$ points ($[-0.24, +5.60]$, $p = 0.09375$). MATH-500 shows a larger high-reuse separation: $+12.81$ AUC points ($[+5.73, +19.89]$, exact $p = 0.03125$) and $+21.55$ points at batch 50 ($[+4.76, +38.34]$, exact $p = 0.0625$). The standalone AIME 2024 gaps at $E = 8$ are $+1.24$ AUC points and $+3.33$ endpoint points; they are compatible with the core-suite pattern but remain descriptive because this benchmark has only 30 problems and was excluded from confirmatory testing. At $E \leq 2$, by contrast, the AUC point estimates favor GRPO on all three benchmarks. The crossover appears at $E = 4$ on the two core benchmarks and at $E = 8$ on all three. These comparisons do not imply that E is a randomized dose of off-policy mismatch: E changes optimizer work, realized policy trajectories, and mismatch simultaneously. As a secondary summary across all four reuse factors, the paired AUC-gap slope per doubling of E is $+0.72$ points on Recent-150 ($[+0.32, +1.12]$, exact $p = 0.0625$) and $+4.44$ points on MATH-500 ($[+2.29, +6.58]$, exact $p = 0.03125$). The corresponding standalone AIME 2024 slope is $+0.41$ points ($[+0.04, +0.78]$) and is descriptive. These linear summaries describe the four realized design cells; they are not causal slopes with respect to mismatch or optimizer count.

The training diagnostics give a compatible but noncausal account. During batches 41–50 at $E = 8$, the GRPO recipe's mean update-averaged $|\log \rho_t|$ is 0.0585 and its oldest-age value is 0.0897, versus 0.0100 and 0.0162 for V-TRACE PO-TC. At $E \leq 4$, every corresponding mean is below 0.006. Figure 7 reports all checkpoint intervals and retains every seed trajectory. The accuracy panels are deliberately unfitted: the observed mismatch is an outcome of training, not an externally assigned treatment.

L.6 COMPLEMENTARY MATCHED-UPDATE ROLLOUT-EFFICIENCY AUDIT

Both the factorial and replicated-curve studies fix the number of fresh rollout batches, so their $E = 8$ arms also receive eight times as many actor updates as an $E = 1$ run of the same length. A separate experiment complements them by matching the actor-update count at its primary comparison. It compares the fixed $\bar{c} = 1, \epsilon = 1, E = 8$ checkpoint after 50 rollout batches and 400 actor updates against six long-horizon $\bar{c} = 0, \epsilon = 0.2, E = 1$ runs evaluated after 100, 200, and 400 rollout batches. At the primary 400-update comparison, both methods receive the same number of actor updates, while the $E = 8$ method uses eight times fewer newly generated responses.

The fail-closed audit covers all six training seeds and all 18 long-horizon $E = 1$ checkpoints on both core endpoints. The fixed $E = 8$ checkpoint averages 9.92% recent-competition pass@1 and 73.50% MATH-500 pass@1. Table 10 compares it with the $E = 1$ learning curves; positive differences favor the original V-TRACE PO-TC recipe.

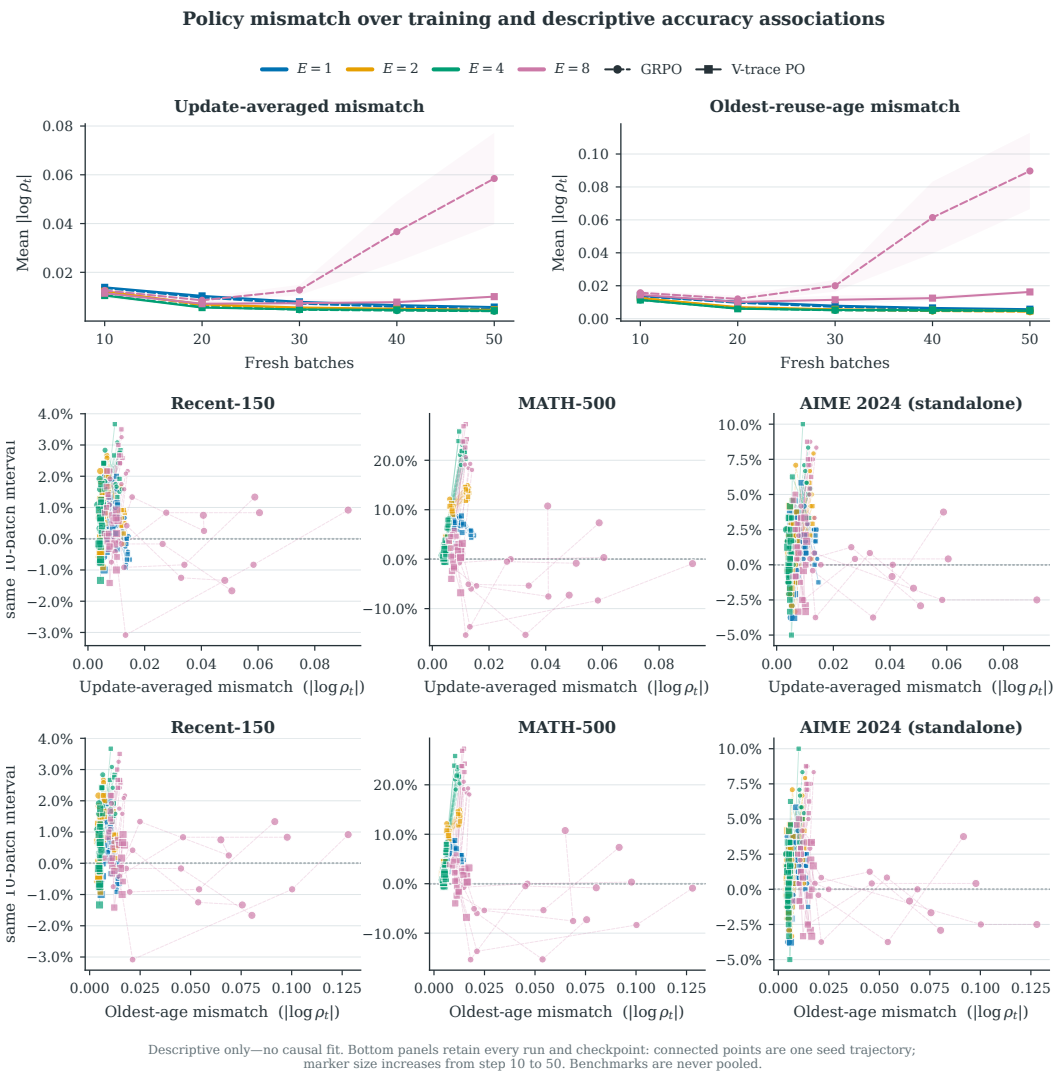

Figure 7: **Realized policy mismatch in the replicated curve study.** Top: mean absolute safeguarded token log-ratio, averaged over all updates or measured at the oldest reuse age, within each ten-batch checkpoint interval; bands are Student-t 95% intervals over six seeds. Bottom: the same interval mismatch against the subsequent ten-batch pass@1 change for every run and checkpoint on each benchmark. Connected points trace a seed, and marker size increases with training time. These panels are descriptive and fit no causal trend: both the optimization recipe and E affect the policy trajectory that generates the horizontal coordinate.

Table 10: Rollout-data efficiency against long-horizon $E = 1$ GRPO. Both methods use Qwen3-8B-Base and MATH training. Recent-competition and MATH-500 sampled pass@1 use eight responses per problem and are reported separately; intervals are paired Student-t 95% intervals over six training seeds.

Endpoint	GRPO batches	GRPO pass@1	Data ratio	Difference	One-sided LB	1-pt NI
Recent-150	100	8.69	2×	$+1.22$ $[+0.14, +2.30]$	$+0.38$	yes, secondary
Recent-150	200	10.61	4×	-0.69 $[-2.62, +1.23]$	-2.20	no
Recent-150	400	11.81	8×	-1.89 $[-3.23, -0.55]$	-2.94	no, primary
MATH-500	100	75.43	2×	-1.93 $[-6.13, +2.28]$	-5.22	no
MATH-500	200	78.40	4×	-4.90 $[-8.59, -1.21]$	-7.79	no
MATH-500	400	81.12	8×	-7.61 $[-10.85, -4.38]$	-10.15	no

At the secondary 100-batch checkpoint, V-TRACE PO-TC minus GRPO is $+1.22$ recent-competition points (95% CI $[0.14, 2.30]$); the one-sided lower bound is $+0.38$, meeting the same one-point tolerance used by the primary noninferiority test. The exact paired sign-flip test is $p = 0.0625$ despite the descriptive Student-t interval excluding zero. This supports half as many generated completions on the recent suite within that tolerance, while using 400 rather than 100 actor updates. It does not extend to MATH-500, where the difference is -1.93 points ($[-6.13, 2.28]$) and the one-sided lower bound is -5.22. Mean training-loop cost is 15.41 rather than 17.48 GPU-hours. This is a tuned-recipe comparison—both $\bar{c}$ and ϵ differ—not an isolated coefficient-cap effect or a claim of lower optimizer work.

At the primary equal-update comparison, V-TRACE PO-TC is lower in all six seeds by 1.89 recent-competition points (95% CI $[-3.23, -0.55]$, exact $p = 0.03125$) and by 7.61 MATH-500 points ($[-10.85, -4.38]$, exact $p = 0.03125$). The preregistered eight-fold noninferiority criterion therefore fails; indeed, the one-sided lower bounds are -2.94 and -10.15, respectively, both below the -1-point margin. At the same 50-batch generation budget, a separate three-seed comparison is $+2.08$ recent-competition points for V-TRACE PO-TC (95% CI $[-4.23, 8.40]$) and $+1.18$ MATH-500 points ($[-11.37, 13.73]$), but V-TRACE PO-TC receives eight times as many actor updates. The evidence supports the directly tested two-fold rollout reduction with additional optimizer work only on the recent suite, not on MATH-500, and does not support four- or eight-fold noninferiority.

Scoped takeaway. The matched-ϵ study provides the cleanest empirical result. On Recent-150, V-TRACE PO becomes more favorable relative to both GRPO and V-TRACE PO-TC from $E = 1$ to $E = 8$; the positive Student-t intervals support reuse robustness, while exact primary p-values of $.0625$ and $.09375$ preclude a 5% rejection. The factorial supplies a distinct mechanism result: under equal data and optimizer work at $E = 8$, the effect of widening the PPO interval depends strongly on the trace-coefficient cap, and the large-cap proxy has extreme raw tails. The older complete-recipe curves triangulate the high-reuse pattern but change both $\bar{c}$ and ϵ. Finally, the matched-update audit does not support lossless four- or eight-fold rollout reduction. Together these studies identify V-TRACE PO as the most reuse-robust arm in the matched setting, without establishing a causal dose response to realized mismatch, universal clipping superiority, or compute efficiency.

M EXPERIMENTAL DETAILS AND MECHANISM STUDIES

M.1 TRAINING, EVALUATION, AND STATISTICAL PROTOCOL

Preregistration, multiplicity and resolution. Each experiment's design, its estimands (the quantities it was built to estimate), its registered predictions and the reading of every possible outcome were frozen and committed before it was launched. The *sequence* of experiments was not: each was chosen in response to what the previous one showed, so the collection as a whole is exploratory even though each element is confirmatory, and no correction is applied across experiments. With six paired seeds the smallest attainable two-sided exact sign-flip p-value is $2/64 = .03125$, so the preregistered tests operate at their resolution floor and no multiplicity correction can be absorbed beneath it; we therefore report effect sizes and intervals alongside every test and treat p-values as descriptive.

All controlled studies start from Qwen3-8B-Base revision `49e3418fbbbc` and train on the MATH training split. A fresh rollout batch contains 32 prompts and eight sampled responses per prompt. Generation is synchronous, uses vLLM with temperature one, top-p one, no top-k restriction, a 1,024-token prompt limit, and a 2,048-token response limit. The policy denominator is the log-probability recorded by the sampler that actually generated each token. Training updates all model parameters with FSDP (LoRA rank zero) in bfloat16. The actor learning rate is 10^{-6}; one PPO minibatch contains all 256 responses from the 32 prompt groups. Each of eight H100 GPUs receives 32 responses and processes them in microbatches of four response rows for gradient accumulation. Training prompts are shuffled using the run's training seed; actor minibatch order is fixed across reuse epochs. Advantages are outcome advantages over the eight responses to each prompt. The earliest experiments used the framework's group-standardized default,

$$A_{gi} := \frac{R_{gi} - \bar{R}_g}{s_g + 10^{-6}}, \tag{56}$$

where g indexes prompts, i indexes the eight sampled responses to prompt g, $\bar{R}_g$ is their mean reward, and s_g is their sample standard deviation. Every experiment from the radius-by-placement factorial onward drops the denominator and uses $A_{gi} := R_{gi} - \bar{R}_g$, so that Corollary C.3 applies exactly; no experiment mixes the two, and Appendix M.14 records the change. Here $R_{gi} \in \{0, 1\}$ is the `math-verify==0.8.0` verdict on the decoded response. Missing or malformed answers, scorer exceptions, and timeouts receive zero. A length-capped response is scored on its retained text rather than automatically assigned zero. An all-equal group has zero advantage. The response-level value is copied to valid response tokens. The policy loss is averaged over valid tokens within each four-response GPU microbatch, and gradient accumulation averages those microbatch objectives. Optimization uses AdamW with $(\beta_1, \beta_2) = (0.9, 0.999)$, weight decay 0.01, a constant 10^{-6} learning rate, and no warmup. Reward-side KL shaping, actor-loss KL regularization, entropy regularization, and reward bonuses are disabled; all runs use norm-one global gradient clipping. Thus at reuse factor E, each fresh batch supplies exactly E consecutive actor updates.

Writing $\delta_t^{\mathrm{fp}} := \mathrm{clip}(\log \pi(a_t \mid x_t) - \log \mu(a_t \mid x_t), -20, 20)$, training code detaches δ_t^{fp} before exponentiation. This safeguarded ratio determines the trace's forward value, while Equation (36) attaches the raw learner score; the clamp is not differentiated. Exact clamp-activation counts were not logged. If the clamp activates, the implemented ratio differs from π/μ, so the change-of-measure identities apply to the mathematical estimator rather than literally to that safeguarded token. Consequently, $\bar{c} = 10^7$ is always described as a large-cap proxy, never as literal untruncated importance sampling. A response that reaches the length cap without an end-of-sequence token is flagged and the reverse recurrence resets at its last retained action. Such a response is useful for optimization under the implementation's finite-horizon convention, but is not evidence for the complete-trajectory identity of Proposition B.6. We therefore report cap-hit rates and do not silently reinterpret truncation as environmental termination.

The core evaluation bundle contains 650 problems and defines two separate suites. The established benchmark suite is canonical MATH-500, containing all 500 problems (Lightman et al., 2024). The recent-competition suite contains 30 problems from each of AIME 2025, AIME 2026, BRUMO 2025, HMMT February 2025, and HMMT November 2025, for 150 problems total; the competition snapshots are distributed through MathArena (Balunović et al., 2025; Dekoninck et al., 2026). Since every recent source contributes the same number of problems, its problem-level micro average equals its equal-source macro average. We report the two suites separately: recent-competition sampled pass@1 is the primary endpoint, while canonical MATH-500 sampled pass@1 is a separately reported key benchmark endpoint. A pooled score over all 650 problems is descriptive only, because MATH-500 supplies most of its weight. Source-specific recent scores, MATH-500 difficulty-stratified scores, and the metadata-defined set of all 105 MATH-500 level-3 problems are secondary diagnostics; no subset is selected using model outcomes.

As a third, standalone endpoint, we report all 30 rows labeled `aime24` in the compiled `ftajwar/srt_test_dataset` release accompanying Shafayat et al. (2025); that paper identifies the source as the public AIME 2024 dataset of Jia (2024). The frozen builder retains source order and performs no outcome-based selection. This familiar benchmark provides a complementary secondary signal, but it is analyzed separately: it enters neither the recent-150 nor MATH-500 score and is omitted from the descriptive pooled-650 score. Because it predates the recent competitions, we make no temporal-separation, post-cutoff, or contamination-free claim for it.

Each frozen target is evaluated with eight responses per problem, temperature one, top-p one, no top-k restriction, repetition penalty one, a 4,096-token response limit, and common evaluation seed 2026. The primary scorer is pinned `math-verify==0.8.0`; malformed answers, scorer exceptions, and timeouts count as incorrect. Sampled pass@1 first averages the eight binary correctness verdicts within a problem and then averages over problems. Observed pass@8 is reported only as a descriptive secondary metric. Raw generations are retained so scorer-only repairs do not resample a model response. Source-specific results, pooled-650 diagnostics, response-cap rates, and seed-level aggregates are reported in the tables below; the pooled score is never used as a headline endpoint. AIME 2024 remains a separate descriptive endpoint.

The five primary competition collections are recent in calendar time, but calendar year alone does not establish that every problem post-dates the undocumented pretraining cutoff of a particular model. We therefore do not label the suite contamination-free. A wrapper- and whitespace-normalized audit that preserves mathematical operators finds no exact prompt overlap among the 650 evaluation problems

or with the MATH training split; this does not exclude semantic overlap or pretraining exposure. AIME 2024 is an older robustness benchmark and is not part of this recency argument.

The suite replaces an earlier benchmark pool for data-eligibility reasons. SMT 2025, CMIMC 2025, HMMT February 2026, and ArXivMath are excluded; AIME 2024 is retained only as the standalone secondary endpoint defined above. Because outcomes on the earlier endpoint were already visible, the replacement is not presented as an untouched confirmation. Dataset contents, model targets, sampling settings, metrics, and analyses were frozen before generating replacement scores, and no earlier benchmark score enters this analysis.

For replicated comparisons, the training seed is the experimental unit. The factorial and matched-update experiments use seeds 2027–2032; the complete-recipe rollout-reuse curves use seeds 2033–2038; and the matched-clipping study uses new seeds 2051–2056. Uncertainty for replicated comparisons is a two-sided Student-t interval across the six relevant seeds, and primary tests use exact paired sign flips. Four prespecified secondary factorial accuracy contrasts on the Recent-150 endpoint share one Holm family: the effect of widening ϵ from .2 to 1 within $\bar{c} = 1$ and within $\bar{c} = 0$; ($\bar{c}$=1, ϵ=1) against the matched baseline ($\bar{c}$=0, ϵ=.2); and ($\bar{c}$=1, ϵ=1) against ($\bar{c}$=10^7, ϵ=1). Their raw exact sign-flip p-values are .03125, .125, .0625 and .375 in that order, giving Holm-adjusted values .125, .25, .1875 and .375; none rejects at familywise 5%. We do not attach inferential intervals to single-seed pilot endpoints: resampling prompts would condition on the realized training seed and cannot supply training-seed replication. Failed numerical runs remain failed outcomes; scorer-only repairs reuse saved generations rather than regenerating model outputs.

The factorial and replicated-curve experiments use frozen training and evaluation configurations. Confirmatory jobs use PyTorch 2.6.0+cu124, Transformers 4.56.1, and vLLM 0.8.5.post1. Before analysis, we verified model and dataset identifiers, licenses, prompt–answer alignment, scorer version, overlap checks, target manifests, run completion records, and all reported figure inputs.

M.2 Exploratory radius sweep and confirmatory lock

The replicated hypothesis was selected from a single-seed transformed-trace PPO sweep crossing $\bar{c} \in \{0, 1, 10^7\}$ with $\epsilon \in \{0.1, 0.2, 0.5, 1\}$ at $E = 8$. Table 11 reports the evaluation-suite endpoints. From $\epsilon = .2$ to 1, recent-competition pass@1 changes by -7.83 points at $\bar{c} = 0$, $+1.42$ points at $\bar{c} = 1$, and $+1.25$ points at the large cap. The corresponding MATH-500 changes are -59.03, $+3.15$, and $+1.55$ points. These are descriptive differences between checkpoints from one training seed, with no training-seed confidence interval. At the wide local endpoint, 93.50% of recent responses and 93.60% of MATH-500 responses hit the 4,096-token evaluation limit, versus 13.92%/5.10% for $\bar{c} = 1$ and 17.00%/7.60% for the large cap.

Table 11: Single-seed exploratory radius sweep at $E = 8$. Entries are sampled pass@1 (percent) with eight responses per problem on the 150-problem recent suite and canonical MATH-500. The $\epsilon = .2$ checkpoints are the shared $E = 8$ anchors from the reuse sweep.

$\bar{c}$	ϵ	Recent-150	MATH-500
0	.1	5.33	43.45
0	.2	8.58	65.75
0	.5	7.25	65.35
0	1	0.75	6.73
1	.1	7.25	67.65
1	.2	7.83	68.68
1	.5	7.83	64.85
1	1	9.25	71.83
10^7	.1	7.25	73.10
10^7	.2	7.83	72.83
10^7	.5	7.83	70.13
10^7	1	9.08	74.38

At $\bar{c} = 0$, mean age-seven $|\delta_t^{\mathrm{fp}}|$ increased from 0.03104 to 0.47662 and mean training reward fell from 0.7318 to 0.3216. At $\bar{c} = 1$, mean $|\delta_t^{\mathrm{fp}}|$ remained between 0.00969 and 0.01350, while the PPO clip fraction fell from 38.62% to 12.50%. These training-side observations supported selecting the

two radii and the interaction estimand; seed 2026 and all exploratory endpoints were excluded from confirmatory inference.

M.3 SEED-LEVEL CONFIRMATORY RESULTS

Table 12 gives the four cells entering the primary recent-competition interaction and the same checkpoints on MATH-500. The recent paired interaction is positive in every seed even though the individual endpoint ordering varies. This is the reason the paper's headline result is an interaction between trace correction and clipping radius, rather than a claim that one checkpoint wins every comparison.

Table 12: Seed-level sampled pass@1 (percent; eight responses per problem) for the confirmatory $E = 8$ study. Panel A is the 150-problem recent primary endpoint; Panel B is canonical MATH-500. D_s is the within-seed difference-in-differences defined in Section L and is computed from unrounded endpoints.

Panel A: recent competitions					
Seed	$\bar{c}$=0, ϵ=.2	$\bar{c}$=0, ϵ=1	$\bar{c}$=1, ϵ=.2	$\bar{c}$=1, ϵ=1	D_s
2027	9.83	4.83	8.58	11.92	+8.33
2028	9.17	5.92	7.42	8.75	+4.58
2029	8.33	1.25	8.00	10.42	+9.50
2030	6.08	6.17	8.75	8.92	+0.08
2031	7.50	8.33	7.83	9.50	+0.83
2032	8.83	2.75	7.25	10.00	+8.83
Panel B: MATH-500					
Seed	$\bar{c}$=0, ϵ=.2	$\bar{c}$=0, ϵ=1	$\bar{c}$=1, ϵ=.2	$\bar{c}$=1, ϵ=1	D_s
2027	65.45	49.98	66.18	75.05	+24.35
2028	63.88	58.10	64.00	67.78	+9.55
2029	55.40	20.30	69.60	75.23	+40.73
2030	60.85	39.75	69.73	69.95	+21.33
2031	60.75	65.40	67.18	76.73	+4.90
2032	61.75	30.73	69.90	76.30	+37.43

The frozen base model scores 5.00% on the recent suite and 43.93% on MATH-500. The complete six-cell means, including the large-cap rows, appear in Table 5, and Figure 4 plots the same audited summaries. Every target has the expected 650×8 samples, common prompt alignment, and zero scorer-error rows.

Table 13: Prespecified secondary contrasts on the recent-competition confirmatory endpoint. Differences and unadjusted Student-t intervals are percentage points; the exact sign-flip tests form one Holm family.

Contrast	Mean	95% CI	Raw p	Holm p
$(\bar{c}=1, \epsilon=1) - (\bar{c}=1, \epsilon=.2)$	+1.94	$[+0.76, +3.13]$	.03125	.125
$(\bar{c}=0, \epsilon=1) - (\bar{c}=0, \epsilon=.2)$	−3.42	$[-6.85, +0.01]$	.125	.250
$(\bar{c}=1, \epsilon=1) - (\bar{c}=0, \epsilon=.2)$	+1.63	$[+0.44, +2.81]$	.0625	.1875
$(\bar{c}=1, \epsilon=1) - (\bar{c}=10^7, \epsilon=1)$	+0.60	$[-0.94, +2.13]$	.375	.375

M.4 POST-HOC SOURCE AND BENCHMARK SENSITIVITY

The primary analysis treats training seed as the experimental unit. Table 14 decomposes the interaction by the five 30-problem sources in the recent suite and reports canonical MATH-500 separately. Every source mean is positive, but three of the five source-specific intervals include zero. These small-source comparisons are descriptive and unadjusted; they do not replace the prespecified aggregate recent-suite test.

Table 14: Six-seed source sensitivity of the $\bar{c} \times \epsilon$ interaction for sampled pass@1 with eight responses per problem. Effects and unadjusted Student-t intervals are percentage points. MATH-500 is a separately reported benchmark, not a constituent of the recent suite. Exact p-values are two-sided paired sign-flip tests and are not part of the primary multiplicity family.

Evaluation	Interaction	95% CI	Exact p
AIME 2025	$+7.78$	$[+1.66, +13.90]$	.0625
AIME 2026	$+4.93$	$[-0.58, +10.44]$	.125
BRUMO 2025	$+10.42$	$[+2.60, +18.23]$	.03125
HMMT February 2025	$+1.53$	$[-0.83, +3.89]$	.21875
HMMT November 2025	$+2.15$	$[-0.49, +4.80]$	.15625
MATH-500	$+23.05$	$[+7.94, +38.15]$	.03125

The MATH-500 interaction is driven by a replicated failure of the wide local update, not by additional data in the corrected arms. The cell means appear in Table 6. Exploratory paired contrasts at $\epsilon = 1$ are $+29.46$ points for $\bar{c} = 1$ versus $\bar{c} = 0$ (95% CI $[+10.41, +48.52]$) and $+28.72$ points for the large cap versus $\bar{c} = 0$ ($[+9.99, +47.46]$); both raw exact sign-flip tests have $p = 0.03125$. The $\bar{c} = 1$ minus large-cap estimate is $+0.74$ points ($[-5.01, +6.50]$, raw $p = 0.75$). Thus MATH-500 reproduces the wide-local-cell failure but provides no accuracy separation between unit truncation and the large cap.

Table 15: MATH-500 sampled pass@1 by official difficulty level with eight responses per problem (percent). Level sizes are shown in the header. Training cells are six-seed means $\pm$ sample standard deviations; the frozen base row is one model evaluation. These metadata-defined strata are descriptive and no level was selected from its outcome.

$\bar{c}$	ϵ	L1 (43)	L2 (90)	L3 (105)	L4 (128)	L5 (134)
Base	—	62.79	57.64	53.10	40.82	24.44
0	.2	76.21 ± 3.13	69.84 ± 4.92	70.95 ± 4.99	60.25 ± 2.95	44.39 ± 3.00
0	1	61.63 ± 19.41	55.30 ± 18.33	51.63 ± 18.70	39.29 ± 17.85	29.43 ± 13.96
1	.2	85.42 ± 2.93	81.06 ± 3.10	79.68 ± 3.16	66.67 ± 2.63	44.87 ± 1.42
1	1	90.31 ± 5.04	85.74 ± 4.64	86.35 ± 4.79	71.57 ± 3.66	51.68 ± 2.40
10^7	.2	87.60 ± 4.43	83.40 ± 4.42	84.52 ± 3.04	69.22 ± 3.91	48.55 ± 3.43
10^7	1	88.37 ± 2.45	85.39 ± 2.76	85.69 ± 2.92	70.36 ± 2.61	51.43 ± 2.98

For the mechanism statistics in Table 5, each run is first averaged over its 50 age-seven observations and then averaged across seeds. For valid token coefficients $g_i = |\operatorname{sg}(\widetilde{\rho}_i)A_i|$, with n valid tokens in the aggregation, the reported normalized coefficient ESS is

$$\mathrm{ESS}(g) := \frac{(\sum_i g_i)^2}{n \sum_i g_i^2},$$

and L1 suppression is $\sum_{i \in \mathcal{C}} g_i / \sum_i g_i$, where $\mathcal{C}$ is the advantage-dependent PPO clipping mask. This measures suppression of the detached scalar coefficient, not bias in the full parameter gradient. In the row order of Table 5, the corresponding mean *training* response-cap fractions are 10.19%, 44.63%, 4.03%, 3.60%, 3.51%, and 3.68%; these are distinct from the 4,096-token evaluation hit-cap rates reported for the single-seed studies.

M.5 Replicated rollout-reuse curve protocol and completeness audit

The rollout-reuse curve experiment is a fully crossed $4 \times 2 \times 6$ design: $E \in \{1, 2, 4, 8\}$, the GRPO recipe ($\bar{c} = 0, \epsilon = .2$) or the original V-Trace PO-TC recipe ($\bar{c} = 1, \epsilon = 1$), and training seed $s \in \{2033, \ldots, 2038\}$. These are 48 distinct training runs. Each run uses the common protocol above for 50 fresh batches (12,800 generated responses), makes $50E$ actor updates, and retains checkpoints at fresh-batch indices $10, 20, 30, 40, 50$. The six seeds are paired across recipe and reuse factor. This pairing does not make the realized rollouts identical after the policies diverge; the inferential unit is the paired training seed.

Every one of the resulting 240 checkpoint targets is evaluated once on the 650-problem evaluation bundle and once on the separate 30-problem AIME 2024 bundle. Eight responses per problem give

exactly 5,200 and 240 generated evaluation rows per target, respectively. The frozen base scores used at curve step zero come from one prior fixed evaluation: 5.00% on Recent-150, 43.93% on MATH-500, and 5.42% on AIME 2024. That point is shared by every displayed curve and is not replicated over training seeds. In particular, its plotted value has no training-seed confidence interval. It cancels from every within-E paired recipe gap because the same value enters both trapezoids.

Let $b_j = 10j$ for $j = 0, \ldots, 5$, and let $Y_{s,m,E}(b_j)$ be sampled pass@1 for seed s, recipe m, and reuse factor E. The normalized curve area used in the primary analysis is

$$\mathrm{AUC}_{s,m,E} := \frac{1}{50} \sum_{j=1}^{5} \frac{b_j - b_{j-1}}{2} \left[Y_{s,m,E}(b_{j-1}) + Y_{s,m,E}(b_j) \right].$$

It has the same units as pass@1. The preregistered Recent-150 interaction is

$$I_s := \left[\mathrm{AUC}_{s,\mathrm{VTP},8} - \mathrm{AUC}_{s,\mathrm{GRPO},8} \right] - \left[\mathrm{AUC}_{s,\mathrm{VTP},1} - \mathrm{AUC}_{s,\mathrm{GRPO},1} \right].$$

Here VTP denotes the original V-TRACE PO-TC recipe. The six values of I_s, in seed order, are $(2.98, 2.45, 3.24, -0.16, 1.77, 2.70)$ percentage points. Their mean is 2.16 points, the Student-t 95% interval is $[0.86, 3.47]$, and the exact two-sided sign-flip value is $p = 0.0625$. The exact result does not reject at 0.05, despite the positive parametric interval. Secondary within-E gaps for every benchmark, including their paired intervals, are reported in Table 9; none converts the recipe comparison into an isolated effect of $\bar{c}$.

Mismatch is summarized without fitting it as an explanatory variable. For each checkpoint, the update-averaged statistic first averages $|\delta_t^{\mathrm{fp}}| = |\operatorname{clip}(\log \pi - \log \mu, -20, 20)|$ across the E reuse ages in each fresh batch and then across the preceding ten batches. The oldest-age statistic instead retains age $E - 1$ before the ten-batch average. At $E = 1$ the two definitions coincide. Table 16 gives the last interval; Figure 7 gives all five intervals and the unfitted accuracy associations.

Table 16: Mean absolute safeguarded token log-ratio during fresh batches 41–50 of the replicated curve study, multiplied by 10^3. "Avg." averages over all reuse ages; "oldest" uses age $E - 1$. Values are means over six training seeds. They are realized outcomes of the two evolving recipes, not randomized mismatch doses.

	GRPO ($\bar{c} = 0, \epsilon = .2$)		V-TRACE PO-TC ($\bar{c} = 1, \epsilon = 1$)	
E	Avg.	Oldest	Avg.	Oldest
1	4.96	4.96	5.76	5.76
2	4.17	4.22	4.61	4.69
4	4.05	4.70	4.28	4.90
8	58.47	89.70	10.03	16.24

The completeness audit verifies all 48 manifests and completion records, exactly 50 logged fresh-batch steps per run, all five checkpoint weight inventories, method-specific hyperparameters, and reuse-age-specific mismatch fields. For evaluation, it requires exactly one complete directory per target and suite; exact prompt–sample grids; model, data, protocol, alignment, and scorer identifiers; and a Boolean verifier result with no scorer error for every row. All 48 training runs, 240 evaluation-bundle evaluations, and 240 AIME evaluations pass.

Infrastructure failures are retained rather than silently erased. The first evaluation arrays stopped before generation when the boxed-scorer checkout did not match the frozen version. Their outputs were quarantined, the checkout was corrected, and one smoke target per suite preceded the full replacement arrays. Five evaluation-bundle tasks later stopped during `math-verify` timeouts after producing 5,088–5,095 of 5,200 rows; each partial directory was quarantined and its exact manifest target was rerun. Six append-only repair-provenance records establish that no model, dataset, sampling protocol, scorer, or analysis choice changed. The analyzer also requires the quarantined inventory to equal the ledger inventory, so a partial run cannot be selected as an outcome-dependent exclusion.

M.6 FROZEN EXECUTION OF THE MATCHED-CLIPPING STUDY

Appendix L.3 defines the estimands. The execution manifest contains exactly 72 unique rows: four reuse factors, the three frozen arms, and six paired seeds. Submission is seed-blocked and

then reuse-blocked, so the three arms for a given (s, E) are adjacent. All arms use synchronous vLLM generation at temperature one, top-p one, top-k disabled, a 2,048-token training-response limit, one 32-prompt minibatch per epoch, per-GPU microbatches of four, learning rate 10^{-6} with a constant schedule, no reward or actor-loss KL term, no entropy bonus, and norm-one global gradient clipping. Generator weights are synchronized before each fresh batch. The rollout-level request seed is intentionally unset because the pinned source would copy it into every repeated sampling request; instead, each paired block sets the data and trainer seeds to its assigned value while retaining the source-default vLLM engine seed.

The launched implementation is fixed to VeRL commit `f4cefba3d42bd771ee0843df1ba3390ba8205090` and experiment commit `e39841ecc320e26229b84ebbaf908fca71b1dfcb` on the dedicated `iclr2027-epochreuse-localclip` branches. Before release, the launcher requires both worktrees to be clean, detached at those published revisions, and verifies the exact Qwen3-8B model-file inventory and SHA-256 digest of every weight, tokenizer, and configuration file. It also records and verifies complete Conda-explicit and `pip freeze --all` inventories, hashes the training and evaluation inputs, and asserts that Python imports VeRL from the pinned source rather than an unrelated editable installation.

The Slurm array is initially held. Its numeric job identifier, design digest, source revisions, input hashes, and environment fingerprints are written atomically before release; any failure before release cancels that exact held array. Each run retains Hugging Face checkpoints after fresh batches 10, 20, 30, 40, and 50. A completion record is written only after every safetensors shard opens successfully and its observed tensor inventory agrees exactly with the checkpoint index. Evaluation is deliberately submitted only after a fail-closed audit of all 72 training outcomes and 360 intended checkpoints. A nonfinite learner outcome remains an algorithmic result; only a genuine infrastructure failure may be repeated, using the identical frozen row and append-only provenance. All 72 training runs and 360 retained checkpoints passed these gates. Evaluation then completed 720 logical targets: 360 on Approved-650 and 360 on standalone AIME 2024, totaling 1,958,400 sampled responses. The final audit selects 712 original outputs and eight protocol-identical Approved-650 repairs (target indices 12, 14, 15, 31, 125, 220, 229, and 315) from repair parent 12261783. The repairs replace failed physical executions of the same frozen targets and do not create additional observations.

M.7 Matched-update comparison: seed-level results

The long-horizon baseline trains $E = 1, \bar{c} = 0, \epsilon = .2$ for 400 fresh rollout batches and retains checkpoints after 100, 200, and 400 actor updates. The comparator is the original V-TRACE PO-TC objective at the fixed $E = 8, \bar{c} = 1, \epsilon = 1$ checkpoint after 50 fresh batches and 400 updates. Table 17 reports every fixed-checkpoint endpoint used by the audit.

On the recent endpoint, V-TRACE PO-TC minus GRPO is $+1.22$ points after 100 GRPO batches ($[+0.14, +2.30]$, raw exact $p = 0.0625$, $p_{\mathrm{Holm}} = 0.125$) and -0.69 points after 200 batches ($[-2.62, +1.23]$, raw and Holm $p = 0.40625$). Their one-sided lower bounds are $+0.38$ and -2.20, so only the two-fold comparison meets the prespecified one-point noninferiority margin. At the primary equal-update comparison, the recent difference is -1.89 points ($[-3.23, -0.55]$, exact $p = 0.03125$; one-sided lower bound -2.94). The corresponding MATH-500 differences after 100, 200, and 400 GRPO batches are -1.93 ($[-6.13, +2.28]$), -4.90 ($[-8.59, -1.21]$), and -7.61 points ($[-10.85, -4.38]$), with raw exact p-values 0.4375, 0.03125, and 0.03125 and one-sided lower bounds -5.22, -7.79, and -10.15; none meets the margin.

All six training jobs reach 400 batches, and all 18 long-horizon evaluation grids contain the expected 650 problems by eight samples with zero scorer-error rows. At 400 updates, the original V-TRACE PO-TC recipe uses 12,800 completions and 16.14 mean full-job GPU-hours, whereas GRPO uses 102,400 completions and 69.46 GPU-hours. The lower resource cost accompanies a 1.89-point recent deficit and a 7.61-point MATH-500 deficit, so it is not a claim of equal-quality eight-fold efficiency.

M.8 Trace cap, PPO radius, and clipping placement

A six-seed 3×2 factorial at $E = 8$ crosses $\bar{c} \in \{0, 1, 10^7\}$ with $\epsilon \in \{.2, 1\}$. The cap-by-radius interaction on Recent-150 is $+5.36$ points ($[0.98, 9.74]$, exact $p = .03125$): widening ϵ changes

Table 17: Seed-level sampled pass@1 (percent; eight responses per problem) in the rollout-efficiency study. GRPO columns use 100, 200, and 400 fresh batches; the original V-TRACE PO-TC recipe uses 50. The two frozen endpoints are reported in separate panels.

Panel A: recent competitions

Seed	GRPO-100	GRPO-200	GRPO-400	V-TRACE PO-TC-50
2027	9.08	10.58	12.58	11.92
2028	8.92	12.00	10.00	8.75
2029	8.83	9.50	10.92	10.42
2030	8.33	11.42	11.25	8.92
2031	8.58	10.00	12.50	9.50
2032	8.42	10.17	13.58	10.00
Mean	8.69	10.61	11.81	9.92

Panel B: MATH-500

Seed	GRPO-100	GRPO-200	GRPO-400	V-TRACE PO-TC-50
2027	75.85	79.83	82.40	75.05
2028	76.25	77.85	78.43	67.78
2029	75.08	78.18	81.10	75.23
2030	75.05	78.13	81.75	69.95
2031	74.98	78.63	83.15	76.73
2032	75.38	77.83	79.88	76.30
Mean	75.43	78.40	81.12	73.50

pass@1 by -3.42 points at $\bar{c} = 0$ but by $+1.94$ at $\bar{c} = 1$. Suffix correction and PPO clipping therefore cannot be tuned independently. Coefficient truncation also controls the raw tail: at reuse age seven the trace p99 is 6.16 at $\bar{c} = 1$ against 1.32×10^{10} at the large cap (Appendix M.13).

This also settles the clipping-placement question. Section 6.1 puts V-TRACE PO ahead of V-TRACE PO-TC at $\epsilon = .2$ by $+1.132$ points at $E = 8$ (exact $p = .03125$), and V-TRACE PO-TC in fact trails plain GRPO at every reuse level. But that radius is not neutral between the two: $\widetilde{\rho}_t$ accumulates suffix deviations and so departs from one far more than ρ_t does, and at $\epsilon = .2$ the transformed-trace mask suppresses 62% of the L1 gradient mass against 6% for local clipping. The two arms were compared at a radius that disables one of them. At $\epsilon = 1$ the placement contrast is $+0.276$ points (exact $p = .66$) and the radius-by-placement interaction is $+0.856$ points ($[-0.407, 2.118]$, exact $p = .15625$). At an appropriate radius the placements are indistinguishable and we make no recommendation between them. An earlier complete-recipe sweep, varying $\bar{c}$ and ϵ together, independently triangulates the high-reuse pattern (Appendix L.5).

M.9 CORRECTION-HORIZON LADDER

The cap family resolves the short-horizon region poorly. Because ρ sits within about 1% of one, $\lambda_t = \min\{1, \bar{c}/\rho_t\}$ moves the horizon from 1 to 2 tokens across $\bar{c} \in [0, 0.5]$ and then jumps to roughly 77 tokens at $\bar{c} = 1$, so the grid $\{0, .5, 1, 2\}$ samples horizons $\{1, 2, 77, \infty\}$ and leaves everything between 2 and 77 unsampled. Under the constant-λ family $c_t = \lambda\rho_t$ we have $\lambda_t = \lambda$ exactly, so $h = 1/(1-\lambda)$ by construction and the horizon can be set directly.

We ran $\lambda \in \{0, .5, .8, .9, .95\}$, giving $h \in \{1, 2, 5, 10, 20\}$, at six paired seeds with $E = 1$ and rollout temperature 1.2: every response is used once, so there is no advantage-sign asymmetry, and the mismatch is the same in every cell. Registered before launch were the shape of Recent-150 curve AUC against h and each horizon against $\lambda = 0$, which is bit-identical to GRPO. Table 18 gives the ladder.

The manipulation check passed: the applied p99 is monotone in λ and $\lambda = 0$ is exactly inert. The registered reading is flat within noise, the spread across arms being smaller than the within-arm seed dispersion. The best arm is $h = 10$ at $+0.060$ Recent-150 points over GRPO, 95% interval $[-0.472, 0.591]$, exact $p = .75$ with four of six pairs positive. The argmax is therefore interior but does not separate from the baseline, so by the registered criterion there is neither an interior

Table 18: **Setting the correction horizon directly leaves accuracy flat.** Six paired seeds at $E = 1$, $T_{\text{gen}} = 1.2$. The applied p99 confirms the horizon is what the construction intends.

λ	horizon h	applied p99	Recent-150 AUC (%)
0	1	1.000	4.639
0.5	2	1.096	4.508
0.8	5	1.210	4.660
0.9	10	1.338	4.699
0.95	20	1.506	4.511

maximum nor a monotone decline. A flat result bounds the size of any short-horizon benefit rather than establishing a decline: it cannot separate "no bias to correct" from "bias exactly offset by variance".

M.10 Coefficient-cap horizon study: cell means

A preregistered ladder in the coefficient cap crossed with two sampler temperatures, six training seeds, 48 runs and 480 evaluations. The learner is held at temperature one, $E = 1$, and the generator is resynchronised every batch, so no self-influence asymmetry exists (cell means in Table 19) and any effect is attributable to the change of measure alone. $T_{\text{gen}} = 1.0$ is a control column carrying only engine nondeterminism.

Table 19: **More correction is never better.** The decline is monotone at $T_{\text{gen}} = 1.0$ and monotone up to noise at $T_{\text{gen}} = 1.2$, where $\bar{c} = 1$ sits 0.006 points above $\bar{c} = 0.5$. Recent-150 normalized curve AUC (%), mean over six seeds with the sample standard deviation, and the applied correction measured as the 99th percentile of the detached multiplier.

$\bar{c}$	0	0.5	1	2
applied w p99, T_{gen}=1.0	1.000	1.067	1.668	2.523
AUC, T_{gen}=1.0	6.256 ± 0.24	6.194 ± 0.41	5.954 ± 0.31	5.576 ± 0.34
applied w p99, T_{gen}=1.2	1.000	1.095	1.859	7.654
AUC, T_{gen}=1.2	4.665 ± 0.30	4.381 ± 0.25	4.387 ± 0.24	4.306 ± 0.14

Realized mismatch is .009 at $T_{\text{gen}} = 1.0$ and .027 at $T_{\text{gen}} = 1.2$. This study uses seeds 2111–2116 and the correction-horizon ladder uses 2121–2126, so their two nominally identical GRPO cells (4.665 and 4.639) differ by seed noise alone. Note that $\bar{c} = 1$, the value used throughout the rest of this paper, is already worse than no correction in both columns, by 0.30 and 0.28 points.

M.11 Advantage-asymmetry probes across regimes

The mechanism probes of Section 4 $\log \mathbb{E}[w \mid A > 0]$ and $\mathbb{E}[w \mid A < 0]$ as detached diagnostics, pooled over every logged update. They are short single-seed probes of roughly ten batches, read for the sign of the effect rather than its size; the same quantity over the full six-seed runs of Appendix M.12 is about six times larger. Table 20 gives the three probes.

Table 20: **The asymmetry tracks reuse, not the size of the mismatch.** Applied advantage-sign gap $\mathbb{E}[w \mid A > 0] - \mathbb{E}[w \mid A < 0]$ and mean per-token $|\log \rho_t|$ in each single-seed probe.

regime	applied gap	mean $\lvert \log \rho_t \rvert$
rollout reuse, $E = 8$	+0.462	.0118
lagging generator	+0.028	.0213
hotter sampler	+0.000	.0368

The ordering is strict and reversed: the regime with the smallest per-token mismatch carries all of the asymmetry, and the hot-sampler probe carries three times the reuse probe's mismatch with none of it. Within the reuse probe the gap also grows with reuse age, from +0.29 at the first update on a fresh

batch to $+0.57$ after five. Reuse age is inferred from logging order, which the age profile depends on and the pooled contrast does not. The nonzero value at age zero, where the mechanism predicts none, is unexplained; it may reflect that offset or a mild correlation between engine-level mismatch and response length.

M.12 Advantage-asymmetry ablation

Both experiments run at $E = 8$, $\epsilon = 1$, with six paired training seeds, 50 fresh MATH batches, mean-only advantages, and the protocol of Appendix M. Throughout, the *applied* multiplier is the factor an arm actually places on the token loss—for the constant arms, the constant itself—logged as a detached diagnostic and pooled over every logged update and every seed. This differs from the proposed multiplier $\widetilde{\rho}_t/\rho_t$, which is invisible to the arms that carry no trace.

The ablation crosses four arms: GRPO; a *fixed pair* replacing the multiplier by constants keyed on $\mathrm{sign}(A)$; a *shuffled* arm that permutes each response's mean multiplier across the microbatch under a derangement, preserving the multiset of magnitudes, every response's length and token shape and all gradient support; and the intact estimator. The shuffle acts inside each response's own mask, so the set of contributing tokens is unchanged. Concretely, each response's own multiplier trajectory is multiplied by a single scalar so that its mean becomes the donor's mean: the within-response token profile, the length and the mask are untouched and no gradient is deleted. An earlier implementation permuted whole rows of the $[B, T]$ multiplier instead; because rows are zero-padded to the longest response in the microbatch, a response paired with a shorter donor received a multiplier of exactly zero on its own valid tokens beyond the donor's length, deleting an estimated 21–33% of gradient tokens and concentrating the loss on the longest responses. That launch was cancelled and none of its output is used. The preregistered manipulation check for the run reported here declared a residual asymmetry of at most ± 0.10; the realized value was $+0.414$ against the intact $+2.857$, so the check failed. The rescaling removed 86% of the asymmetry rather than all of it, because it can only decorrelate within a microbatch and roughly an eighth of microbatches carry a single advantage sign. The residual asymmetry therefore inflates *shuffled minus GRPO*, so that contrast bounds from above how much benefit survives without the coupling; *intact minus shuffled* is confounded in both directions, since the residual deflates it while the added noise inflates it. The fixed-pair and intact contrasts against GRPO contain no shuffled data and are unaffected.

How the shuffled arm is to be read. The permutation attaches each response's magnitude to a different response, so it adds noise as well as removing structure, and the preregistration said so before the run: a decline of the shuffled arm relative to the intact estimator is consistent either with the asymmetry mattering or with the added noise. The discriminating contrast is therefore *shuffled minus GRPO*, not *intact minus shuffled*. Two readings were registered in advance: if the shuffled arm lands at GRPO's level the asymmetry account is supported, and if it lands below both the ablation was too destructive to interpret and should be replaced by two-copy cross-fitting. It landed $+0.936$ above GRPO, which is neither — consistent with the 86% removal recorded above rather than with a complete ablation. One further comparison bears on the noise term: the fixed pair and the shuffled arm apply almost the same asymmetry (0.408 and 0.414) by entirely different constructions, yet recover $+1.387$ and $+0.936$ respectively. A gap between two arms at matched applied asymmetry is what the declared noise cost would look like, though six seeds cannot establish that it is the cause. Table 21 gives all five contrasts.

Table 21: Ablation: registered Recent-150 AUC contrasts, in percentage points, with 95% Student-t intervals and exact two-sided sign-flip p-values over six paired seeds.

contrast	mean	95% CI	p	+pairs
V-TRACE PO − GRPO	$+2.682$	$[1.637, 3.727]$	.03125	6/6
V-TRACE PO − shuffled	$+1.746$	$[-0.102, 3.594]$	.125	4/6
V-TRACE PO − fixed pair	$+1.294$	$[-0.668, 3.257]$	.15625	4/6
fixed pair − GRPO	$+1.387$	$[0.211, 2.564]$	.0625	5/6
shuffled − GRPO	$+0.936$	$[-0.161, 2.033]$	.09375	5/6

A dose-matched follow-up was run at the same cell and seeds, replacing the multiplier by the trace's own measured levels. It bears on asymmetric reweighting rather than on suffix correction — it sets a

tuned pair against an untuned algorithm, and the trace's realized gap spans $+2.03$ to $+7.81$ across these six seeds, so no constant is matched to any particular run. It is therefore out of scope here and is not used to support any claim in this paper.

M.13 Trace cap and PPO radius: supporting detail

Material supporting Appendix M.8.

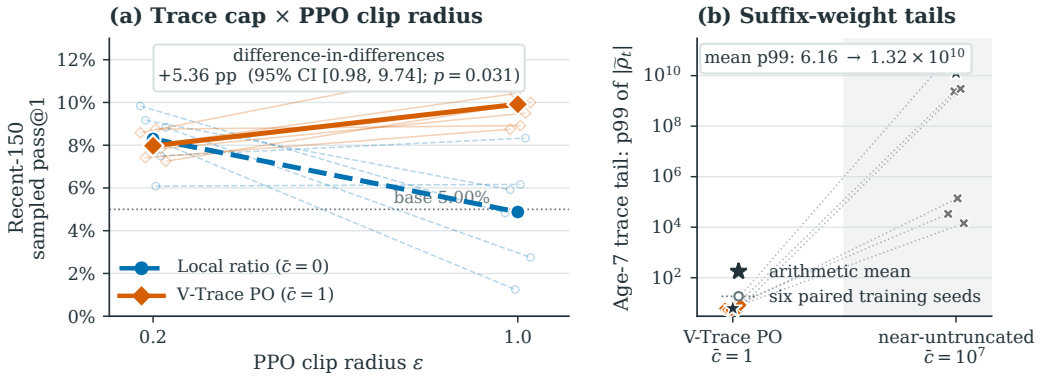

Figure 8: **The trace cap changes the usable PPO radius.** Left: paired Recent-150 endpoints at $E = 8$. Right: reuse-age-seven trace p99; the large-cap proxy has a much heavier raw tail.

Figure 8 shows the paired endpoints and the raw trace tails. The wide local endpoint at $\bar{c} = 0$, $\epsilon = 1$ is degenerate: 93.5% of Recent-150 responses hit the evaluation cap. Because the primary interaction subtracts that arm, part of its magnitude measures how badly this cell fails rather than how much the trace helps — the same qualification the main text makes for the radius-by-placement factorial. Without the outer PPO mask, all three large-cap stress runs encounter nonfinite diagnostics, whereas every unit-cap run remains finite. Full factorial results are in Table 5.

On the placement question, the quantities behind the main-text summary are as follows. Even with a freshly synchronised sampler the trace p99 is 1.80, and the counterfactual trace-clip rate is 6.8% against 0.06% for the local mask, which is why the $\epsilon = .2$ comparison is not neutral between the placements. The per-doubling placement slope falls from $+0.273$ points at $\epsilon = .2$ (exact $p = .0625$) to $+0.111$ at $\epsilon = 1$ (exact $p = .469$), and the radius-by-placement interaction at $E = 8$ is positive in five of six seed pairs.

The earlier complete-recipe sweep compares standard GRPO ($\bar{c} = 0, \epsilon = .2$) with the original V-Trace PO-TC recipe ($\bar{c} = 1, \epsilon = 1$). Its Recent-150 AUC gap shifts from -0.54 points at $E = 1$ to $+1.63$ at $E = 8$, an interaction of $+2.16$ points ($[0.86, 3.47]$, exact $p = .0625$). It triangulates the high-reuse pattern but does not isolate suffix correction or clipping placement, because both $\bar{c}$ and ϵ change together; Section L.5 retains the full curves, aggregate table, and mismatch diagnostics. The complementary equal-update audit in Appendix L.6 does not support lossless eight-fold rollout replacement.

M.14 Radius-by-placement factorial: cell means

This preregistered experiment completes the matched-clipping design of Section 6.1 at a second clipping radius: three arms crossed with $\epsilon \in \{.2, 1\}$, $E \in \{1, 2, 4, 8\}$ and six training seeds, for 144 runs and 1,440 evaluations over 720 targets. Cell means are in Table 22, and Figure 9 follows the $E = 8$ cells over training.

Why the $\epsilon = .2$ cells differ from Table 8. This is a separate experiment, not a re-scoring of the earlier one, and its $\epsilon = .2$ half is a fresh set of 144 training runs rather than the earlier 72. Two things differ by design. It uses training seeds 2081–2086 where the matched-clipping study used 2051–2056, so no cell is paired across the two. And it uses *mean-only* advantages where the earlier study used the framework's group-standardized default, a protocol change adopted from this experiment onward so

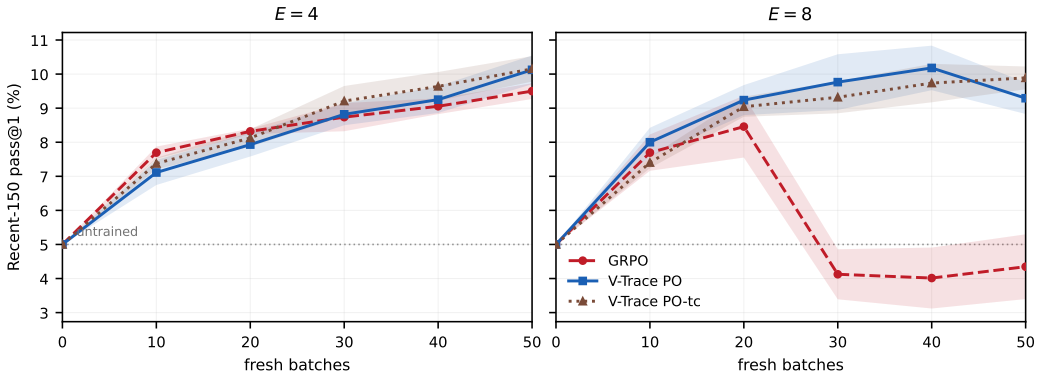

Figure 9: **What the last doubling of reuse does.** Recent-150 pass@1 against fresh-batch index at $\epsilon = 1$, mean over the six training seeds with the standard error of the mean; the dotted line is the untrained model. At $E = 4$ all three arms train normally and finish together. At $E = 8$ matched GRPO rises to 8.46 by batch 20 and then falls to 4.12, below where it started, while both trace placements continue to climb; V-TRACE PO peaks at 10.18, the highest point reached by any arm at this radius. This is the cell that supplies the $+3.07$ headline gap, and the figure shows it is not only GRPO's collapse: V-TRACE PO is also better here than at any smaller reuse depth.

that Corollary C.3 applies exactly (Section 3). The two tables should therefore be read as independent estimates of the same qualitative effect under different advantage normalization, and the agreement of their per-doubling slopes ($+0.298$ and $+0.328$ points) is a replication across that change rather than a consistency check on one dataset. Neither is a re-scoring of the other; the 93 scorer hard timeouts noted above are internal to this experiment and were scored incorrect under the frozen protocol. All other settings match the protocol of Section 6.

Table 22: **Radius-by-placement factorial, Recent-150 normalized curve AUC (%).** Mean over six training seeds with the sample standard deviation. The reuse benefit of V-TRACE PO over GRPO grows with E at both radii and is largest at $E = 8$, $\epsilon = 1$; the two clipping placements are separated at $\epsilon = .2$ but not at $\epsilon = 1$.

ϵ	E	GRPO	V-TRACE PO	V-TRACE PO-TC
.2	1	6.299 ± 0.26	6.319 ± 0.21	5.856 ± 0.23
.2	2	7.301 ± 0.23	7.424 ± 0.22	6.904 ± 0.19
.2	4	8.229 ± 0.48	8.421 ± 0.28	7.179 ± 0.19
.2	8	6.939 ± 0.96	8.031 ± 0.57	6.899 ± 0.40
1	1	6.189 ± 0.27	5.947 ± 0.29	6.115 ± 0.17
1	2	7.385 ± 0.43	7.257 ± 0.37	7.286 ± 0.26
1	4	8.211 ± 0.38	8.135 ± 0.57	8.385 ± 0.57
1	8	$\mathbf{5.793 \pm 1.45}$	$\mathbf{8.865 \pm 1.01}$	8.589 ± 0.65

Per-doubling slopes of the paired AUC gap, with exact two-sided sign-flip p-values over the six seeds: Local minus GRPO is $+0.328$ points $[0.046, 0.610]$ at $\epsilon = .2$ and $+0.999$ points $[0.177, 1.821]$ at $\epsilon = 1$ (both $p = .0625$); Local minus Trace is $+0.273$ points $[0.039, 0.506]$ at $\epsilon = .2$ ($p = .0625$) and $+0.111$ points $[-0.291, 0.514]$ at $\epsilon = 1$ ($p = .469$). The radius-by-placement interaction on the $E = 8$ contrast is $+0.856$ points $[-0.407, 2.118]$, exact $p = .15625$, with five of six pairs positive: widening the radius gains V-TRACE PO-TC $+1.690$ points and V-TRACE PO $+0.835$.

M.15 CLIPPING-PLACEMENT PILOT

Before the replicated factorial, a single training seed compared the three objectives in Table 3 at $E = 8$ and $\epsilon = 0.2$. Every cell used 50 fresh batches and 400 actor updates unless the numerical stopping rule fired. At $\bar{c} = 0$, transformed-trace PPO and local-ratio PPO are algebraically identical;

deterministic value, gradient, and clip-fraction tests verified this equality, so the common control was trained once.

Table 23: Complete single-seed clipping-placement matrix. Trace p99 is the mean of logged microbatch p99 summaries at reuse age seven; gradient maximum is the largest finite pre-global-clipping norm at that age. Endpoints are sampled pass@1 with eight responses per problem on the 150-problem recent suite and canonical MATH-500. These descriptive diagnostics condition on one trained checkpoint per cell and carry no training-seed confidence interval.

Objective	$\bar{c}$	Batches	Trace p99	Grad. max	Recent-150	MATH-500
V-TRACE PO-TC	0	50	1.225	5.66	8.58	65.75
V-TRACE PO-TC	1	50	2.557	.203	7.83	68.68
V-TRACE PO-TC	10^7	50	17.617	.149	7.83	72.83
V-TRACE PO	1	50	10.049	90.9	7.33	71.25
V-TRACE PO	10^7	50	7.08×10^4	1.93×10^{14}	10.33	76.20
V-TRACE PG	0	50	2.515	1.71×10^4	2.58	42.70
V-TRACE PG	1	50	523.408	1.56×10^5	6.58	55.00
V-TRACE PG	10^7	41, failed	1.84×10^{24}	1.16×10^{19}	—	—

The transformed-trace clip masks the suffix weight itself: at reuse age seven it clips 38.62% of tokens for $\bar{c} = 1$ and 39.60% for the large cap, keeping the largest logged gradient norm below 0.21. Under local-ratio clipping, only 0.70% and 0.35% of local ratios are clipped; the detached continuation multiplier remains outside that mask, producing the 10^{14}-scale large-cap gradient in Table 23. Removing the PPO mask entirely is more severe: the large-cap run records 1,117 nonfinite token weights during its 42nd attempted batch and exits without a final checkpoint. All arms still share norm-one global gradient clipping, so completion of a finite run is not evidence that its raw estimator is well conditioned.

The endpoint columns deliberately accompany the stability statistics. They do not rank clipping placements uniformly: at $\bar{c} = 1$, transformed-trace clipping is 0.50 points higher on the recent suite, whereas local clipping is 2.58 points higher on MATH-500. The local-clipping large-cap checkpoint is numerically highest on both endpoints, but it is one seed and its logged gradient maximum is 1.93×10^{14}. The stable conclusion from this pilot is therefore mechanistic: putting a long suffix multiplier outside the PPO mask exposes the optimizer to its full tail, while transformed-trace clipping subjects the cumulative weight to that mask. The six-seed factorial kept transformed-trace PPO fixed and tested whether coefficient-capped correction changes its useful clipping range. It did not rank clipping placements. The 72-run study in Appendix L.3 provides that replicated comparison at $\bar{c} = 1$ and $\epsilon = .2$.

M.16 EXPLORATORY SINGLE-SEED REUSE-SWEEP PRECURSOR

Before the replicated recipe curves, a one-seed exploratory sweep crossed $E \in \{1, 2, 4, 8\}$ with $\bar{c} \in \{0, 1, 10^7\}$ at $\epsilon = 0.2$, using the original transformed-trace clipping placement in every arm. It fixes the number of fresh rollout batches at 50, so increasing E also increases actor updates from 50 to 400. Table 24 reports fixed-checkpoint endpoints on the recent suite and separately reported MATH-500. Because it uses one seed and holds the narrow radius fixed, it neither supplies replication nor compares the two complete recipes in Section L.5.

The paired correction-minus-local-ratio contrasts are:

The recent-suite point estimates do not show a correction advantage at $E \in \{2, 4, 8\}$; at $E = 8$, the local-ratio arm scores 8.58% and both corrected arms score 7.83%. MATH-500 tells a different single-seed story: at $E = 8$, the local-ratio arm scores 65.75%, versus 68.68% for the unit-capped trace and 72.83% for the large cap. The corresponding 4,096-token evaluation-cap rates are 23.00%, 7.55%, and 5.43% on MATH-500 and 29.25%, 13.25%, and 8.83% on the recent suite. This benchmark-sensitive failure in one training seed motivated the replicated clipping-radius experiment and the later recipe-curve study; it is an exploratory precursor rather than headline evidence.

The realized mismatch documents why $E = 8$ became the informative stress regime in this precursor. At the oldest reuse age, mean $|\delta_t^{\mathrm{fp}}|$ for $(\bar{c} = 0, 1, 10^7)$ is respectively $(.0083, .0095, .0099)$ at $E = 1$,

Table 24: Exploratory single-seed rollout-reuse precursor: sampled pass@1 (percent) with eight responses per problem. The first three columns are the 150-problem recent suite and the last three are canonical MATH-500. Cross-E comparisons also change the number of optimizer updates; all values are descriptive.

	Recent-150 pass@1			MATH-500 pass@1		
E	$\bar{c}=0$	$\bar{c}=1$	$\bar{c}=10^7$	$\bar{c}=0$	$\bar{c}=1$	$\bar{c}=10^7$
1	6.92	7.33	7.33	71.85	67.78	64.33
2	8.83	8.75	8.42	74.13	73.03	70.45
4	9.50	7.92	8.33	76.05	73.48	71.68
8	8.58	7.83	7.83	65.75	68.68	72.83

Table 25: Within-E correction-minus-local-ratio pass@1 contrasts in percentage points. The realized checkpoints share one training seed, so these descriptive differences have no training-seed confidence interval.

E	Recent-150		MATH-500	
	$\bar{c}=1$ minus 0	$\bar{c}=10^7$ minus 0	$\bar{c}=1$ minus 0	$\bar{c}=10^7$ minus 0
1	$+0.42$	$+0.42$	-4.08	-7.53
2	-0.08	-0.42	-1.10	-3.68
4	-1.58	-1.17	-2.58	-4.38
8	-0.75	-0.75	$+2.93$	$+7.08$

$(.0068, .0073, .0079)$ at $E=2$, $(.0062, .0072, .0075)$ at $E=4$, and $(.0310, .0097, .0084)$ at $E=8$. The values are outcomes of independently evolving policies, not a monotone dose variable; they show both a nonzero sampler–trainer mismatch floor at age zero and the additional drift that can emerge after repeated optimization of one rollout batch.

M.17 Numerical safeguards and reporting conventions

The implementation checks every trace and transformed coefficient for finite values before optimization and terminates a run when the check fails. It never silently drops nonfinite tokens. Valid-token masks exclude padding, and the reverse recurrence resets at response and packed-segment boundaries. The division-free construction in Equation (36) avoids dividing by an underflowed local ratio in the transformed-trace branch; the default V-TRACE PO branch instead forms the multiplier from $w_t = 1 + \lambda_t(\widetilde{\rho}_{t+1} - 1)$, which needs no division either. Global norm clipping is applied only after backward propagation, so it can bound a large finite aggregate gradient but cannot rescue a nonfinite trace or loss.

The setting $\epsilon = 1$ is still a clipped PPO objective with nominal range $[0, 2]$ and an advantage-dependent derivative mask; it is not V-TRACE PG. Within each microbatch, trace p99 is the empirical 99th percentile of $|\widetilde{\rho}_t|$ over valid response tokens. Reported p99 values average these microbatch quantiles over updates (and seeds where applicable); they are not recomputed from one pooled token population. We therefore interpret their replicated orders of magnitude rather than small numerical differences.