T^5: TWIN-CRITIC TRAINING FOR TOKEN-LEVEL THOUGHTS IN REINFORCEMENT MID-TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement mid-training lets language models learn internal thoughts from unlabeled text, but efficient token-level credit assignment remains challenging. Existing group-relative methods require costly repeated generation. Learned critics offer single-rollout feedback, but accurate return prediction alone does not ensure reliable policy updates. Our analysis shows how training–inference mismatch and PPO clipping prevent a common offset in advantage estimates from cancelling out, introducing additional update drift. We propose T^5, a twin-critic method that calibrates token-level advantages from a single generated trajectory. After warmup and held-out qualification, the critics provide two advantage estimates, combined using action-dependent weights learned through a conditional-moment saddle-point objective. This objective brings the average advantage at each prefix toward zero, while a signal-retention constraint prevents the correction from erasing the learning signal. Sharing information across text positions avoids repeated sampling of each prefix. Theoretically, we characterize optimal mixing under the signal-retention constraint and establish an upper bound on residual mean-induced drift. Experiments show that, compared with the state-of-the-art critic-free method, T^5 improves mean benchmark performance by 7.8% and reduces mean training-step time by up to 63.4%.

1 INTRODUCTION

Language models can learn to reason from ordinary text by generating internal thoughts that help predict what comes next. In QUIET-STAR, improved continuation likelihood rewards these thoughts without labeled answers or reasoning traces (Zelikman et al., 2024). Recent work brings reinforcement learning into pre- and mid-training: RPT and RLPT turn next-token and next-segment prediction into reasoning tasks (Dong et al., 2025; Li et al., 2025). RLP rewards the information gained from thoughts, while PretrainZero learns to select and predict masked spans (Hatamizadeh et al., 2026; Xing et al., 2025). RMT targets the intermediate training stage with adaptive thought-length limits and curriculum sampling (Tian et al., 2025). A central challenge remains token-level credit assignment: a useful continuation does not reveal which thought tokens contributed.

Providing this token-level feedback, however, can require substantial generation. Critic-free group-relative methods such as GRPO use multiple responses to construct a baseline (Shao et al., 2024), as do RMT and RLP (Tian et al., 2025; Hatamizadeh et al., 2026). We use groups of eight throughout our GRPO comparisons. A learned critic instead shares return predictions across texts and prefixes, as explored in value-based LLM optimization (Yue et al., 2025). By predicting the remaining return from a partial thought, it can combine prefix rewards with generalized advantage estimation (GAE) to provide token-level feedback along a single trajectory (Schulman et al., 2016; 2017).

Efficient single-rollout training also has to accommodate discrepancies between generation and optimization. Reusing rollouts introduces policy lag (Schulman et al., 2017), while differences in numerical precision, kernels, and batching can make training and inference probabilities disagree even at the same model parameters (Marek & Ryabinin, 2026). Stabilization methods address different aspects of these discrepancies: GSPO uses sequence-level reweighting (Zheng et al., 2025), CPPO constrains position-dependent and accumulated prefix drift (Mao et al., 2026), and Score Centering corrects the biased expectation of the policy score—the log-probability gradient—under sampling mismatch (Marek & Ryabinin, 2026). These methods regulate sampling and policy updates. A

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.

learned critic must also predict current returns reliably. Newly initialized value heads therefore need warmup and checks on current, held-out returns before guiding actor updates.

However, reliable return predictions alone do not remove the effect of training–inference mismatch on advantage-based updates. At a given thought prefix, advantage estimates may rank actions correctly yet be uniformly too high or too low. This common offset cancels from the expected policy gradient under matched sampling without clipping, but mismatch or PPO's sign-dependent clipping can prevent that cancellation. Our analysis separates the expected update into a component reflecting variation across actions and a mean-induced drift component. The latter couples the advantage offset to the average update direction after importance weighting and clipping. This identifies an advantage-side complement to the sampling and score corrections above: bring the prefix-conditioned advantage mean toward zero without erasing the learning signal.

To control this mean-induced drift, we propose T^5, a twin-critic method that learns how to combine two advantage estimates. After warmup and held-out qualification, both critics evaluate the same generated thought. An action-dependent weight mixes their estimates at each token. We learn this weight through a conditional-moment saddle-point objective: a maximizing auxiliary mean predictor exposes the mixture's residual offset, while the minimizing weight uses critic disagreement to reduce it. The mean predictor learns across contexts, using one thought per selected position for both estimates rather than repeated prefix continuations. A constraint limits deviation from the fixed twin average to preserve nonzero aggregate signal. This saddle-point formulation links calibration to optimal mixing and a bound on residual mean-induced drift.

Our contributions are this drift analysis, a prediction-qualified single-rollout calibration method, and its theoretical guarantees. In the population formulation, we prove strong duality and characterize optimal mixing under the signal-retention constraint. We also bound retained signal strength and residual mean-induced drift, accounting for mean-prediction error and validation uncertainty. Compared with RLP, the state-of-the-art critic-free reinforcement pretraining method, T^5 improves mean benchmark performance by 7.8% and reduces mean training-step time by up to 63.4%.

2 Related Work

Latent reasoning and reinforcement mid-training. Quiet-STaR learns hidden rationales from their effect on future-token likelihood (Zelikman et al., 2024), while Fast Quiet-STaR compresses explicit thought tokens (Huang et al., 2025). These methods connect intermediate computation to ordinary text, complementing reasoning-trace bootstrapping such as STaR (Zelikman et al., 2022). Reinforcement Pre-Training, RL on Pre-Training Data, and RLP also construct reinforcement signals from pretraining corpora (Dong et al., 2025; Li et al., 2025; Hatamizadeh et al., 2026). PretrainZero selects masked spans for self-supervised reinforcement pretraining (Xing et al., 2025). RMT combines reinforcement mid-training with thought-token allocation, curriculum sampling, and next-token prediction (Tian et al., 2025). Our task follows Quiet-STaR's continuation-utility objective at this intermediate training stage. The focus is whether learned token-value baselines can supply reliable PPO advantages across the resulting thought prefixes.

Group-relative estimation and token-level actor–critic learning. GRPO replaces the learned value baseline with comparisons among responses sampled for a prompt (Shao et al., 2024), a route also used for large-scale reasoning reinforcement learning (DeepSeek-AI, 2025). A learned critic instead shares information across sampled states and supports GAE along each trajectory (Schulman et al., 2016; 2017). VAPO studies value-based LLM optimization and introduces the length-adaptive trace used here (Yue et al., 2025). SAO learns baselines under single-rollout asynchronous agentic training (Hou et al., 2026). Related offline-RL work addresses critic-side instability by controlling harmful TD cross-covariance or modifying optimizer dynamics to suppress critic collapse (Qiao et al., 2026a;b). Our states are hidden thought prefixes and rewards come from continuation prediction, with a fixed rollout law within each window. Twin critics in TD3 and SAC manage function-approximation error through action-value targets (Fujimoto et al., 2018; Haarnoja et al., 2018). Our two state-value critics fit observed returns, and a conditional-moment objective learns how to combine their advantages.

What policy stabilization controls. Existing methods act on different parts of the update. TRPO constrains policy displacement and PPO clips probability ratios (Schulman et al., 2015; 2017). REINFORCE++ and Dr. GRPO change advantage or loss normalization (Hu et al., 2025; Liu et al., 2025). DAPO combines clipping and sampling changes with token-level loss aggregation (Yu et al., 2025). GSPO uses sequence-level ratios, while CPPO makes the trust region position- and prefix-dependent (Zheng et al., 2025; Mao et al., 2026). Score Centering addresses update drift under training–inference engine mismatch through an additive score correction (Marek & Ryabinin, 2026). T^5 retains clipped PPO and acts on the conditional advantage mean, after qualifying the predictors that produce it. This distinction respects the usual baseline-cancellation result: inaccurate state baselines can still cancel under exact unclipped score weighting (Williams, 1992; Sutton et al., 2000). The issue here is their interaction with bootstrapping and incomplete score cancellation.

3 Preliminaries and Problem Setup

3.1 Reinforcement Mid-Training from Unlabeled Text

Reinforcement mid-training uses ordinary text to supervise useful internal computation before task-specific post-training (Tian et al., 2025; Hatamizadeh et al., 2026). From a Base model, we sample a T-token hidden thought $z_{1:T}$ at position p in a length-S text $x_{1:S}$ drawn from corpus $\mathcal{D}_{\mathrm{mid}}$. Following QUIET-STAR (Zelikman et al., 2024), we score the thought by its effect on prediction loss over the next H observed text tokens. After t thought tokens, this loss is

$$\ell_t = -\frac{1}{H}\sum_{j=1}^{H} \log p_{\mathrm{score}}(x_{p+j} \mid x_{\le p}, z_{1:t}, x_{p+1:p+j-1}), \tag{1}$$

where p_{score} is the scoring model's token distribution, and ℓ_0 is the loss without a thought. The full thought has utility $\mathcal{U} = \ell_0 - \ell_T$: it compares prediction of the same continuation with and without the thought. Each observed token contributes a log-likelihood gain. The changes after successive thought tokens also add up to $\mathcal{U} = \sum_{t=1}^{T}(\ell_{t-1} - \ell_t)$. An individual change may be negative even when the full thought helps.

The actor samples token $a_t = z_t$ from the visible prefix $s_t = (x_{\le p}, z_{<t})$. Actor and critic inputs exclude the scoring text. To assign rewards along the thought, write $\Psi_t = \ell_0 - \ell_t$ for the gain after its first t tokens. We scale and clip this gain, then reward each change in the resulting potential:

$$\Phi_t = \mathrm{clip}(\Psi_t/\sigma_r, -c_r, c_r), \qquad r_t = \Phi_t - \Phi_{t-1}, \quad \Phi_0 = 0, \tag{2}$$

where $\sigma_r > 0$ is the reward scale and $c_r > 0$ the clipping threshold. Between scoring checkpoints we carry the last potential forward, so $\sum_t r_t = \Phi_T = \mathrm{clip}(\mathcal{U}/\sigma_r, -c_r, c_r)$. The scorer and reward scale stay frozen within each training window (Appendix E.1). For analysis, c denotes the full pre-action context, including the visible prefix s, scoring text, and checkpoint history. Unless stated otherwise, conditioning on s also fixes these other parts of c. This shorthand does not change what the actor and critic observe.

Training and prediction. Scoring and critic evaluation occur during training. A gate $\beta_g(s) \in [0, 1]$ combines thought-conditioned and no-thought predictions, defining the mixed gain in Figure 1. Ordinary next-token loss anchors language modeling. This prediction gate is separate from the advantage weight in Section 5. Its objective detaches expert logits and features (Appendix E.5). Scorer and generation settings are specified with the evaluation protocol (Appendix A).

3.2 PPO with a Learned Token-Level Critic

A learned critic shares return information across texts and partial thoughts. PPO combines this reusable baseline with the prefix rewards above, which can also support critic-free estimators. Let π_θ be the actor with parameters θ, π_{old} the recorded rollout policy, and q the actual sampling law. A critic $V_\psi(s_t)$, with parameters ψ, predicts remaining return $G_t = \sum_{k=0}^{T-t} \gamma^k r_{t+k}$ with discount γ. Generalized advantage estimation (GAE) gives (Schulman et al., 2016)

$$\delta_t = r_t + \gamma V_\psi(s_{t+1}) - V_\psi(s_t), \qquad A_t = \delta_t + \gamma\lambda A_{t+1}, \tag{3}$$

where δ_t is the temporal-difference residual, A_t the advantage estimate, and λ the trace parameter controlling how far later residuals propagate. We use $\gamma = 1$, full-return targets, and a length-adaptive trace (Appendix E.3). Values and advantages vanish after termination.

With $\widetilde{A}_t$ denoting a normalized, detached advantage, PPO maximizes (Schulman et al., 2017)

$$\mathcal{J}_{\text{clip}}(\theta) = \mathbb{E}_{\mathcal{D}}\Big[\min\{\rho_t \widetilde{A}_t, \text{clip}(\rho_t, \rho_-, \rho_+)\widetilde{A}_t\}\Big], \tag{4}$$

where $\rho_t = \pi_\theta(a_t \mid s_t)/\pi_{\text{old}}(a_t \mid s_t)$ is the token probability ratio and $0 < \rho_- \leq 1 \leq \rho_+$ are its lower and upper clipping bounds, which may be asymmetric. Replay $\mathcal{D}$ averages valid actor actions. Recorded probabilities, critic snapshots, and advantages stay fixed during actor updates.

Unlike the learned predictor V_ψ, $V^q(s) = \mathbb{E}_q[G_t \mid s_t = s]$ denotes the conditional expected return under q, and V^{π_θ} is its current-actor counterpart. Policy lag is movement from π_{old} to π_θ, while sampling mismatch is a difference between q and π_{old} (Marek & Ryabinin, 2026). Critic mismatch concerns prediction relative to V^q, including stale fitting targets or hidden scoring information (Appendix D.2). Refreshing the rollout copy does not refit the critic. GRPO instead uses current group rewards, avoiding learned-value history but retaining policy lag and finite-group estimation error.

Statistical notation. For any scalar or vector quantity X, write $\mu_X(s) = \mathbb{E}_q[X \mid s]$ for its conditional mean. The symbols Var_q and Cov_q denote variance and covariance. We omit token index t when discussing a generic prefix. The finite measure d weights rollout prefixes, and $\mathbb{E}_d$ denotes the resulting aggregation, while $\mathbb{E}_{d,q}$ additionally averages sampled actions and continuations. These weights need not sum to one, but their scale is fixed across the learning objectives (Appendix E.4).

4 Critic Mismatch and Predictive Qualification

4.1 The Critic Adds Its Own Mismatch

Clipping constrains how far the actor moves from its recorded policy, but it does not tell us whether the value reference still describes current continuation returns. Keeping the critic aligned with a changing policy is a longstanding actor–critic concern (Konda & Tsitsiklis, 2003; Yue et al., 2025). Let q_β denote the earlier rollout law that supplied the critic's fitting data and q the current rollout law. With the scoring rule fixed,

$$V_\psi - V^{\pi_\theta} = \underbrace{(V_\psi - V^{q_\beta}) + (V^{q_\beta} - V^q)}_{\text{fit error + target lag}} + \underbrace{(V^q - V^{\pi_\theta})}_{\text{rollout–actor gap}}.$$

The first two terms arise from fitting and reusing a learned critic. Even under an idealized refresh with $q = \pi_{\text{old}} = \pi_\theta$ and $\rho = 1$, the rollout–actor gap vanishes while these critic-specific terms can remain. GRPO avoids this learned-critic history by constructing its baseline from the sampled group (Shao et al., 2024). Critic-based PPO therefore needs value tracking in addition to policy-ratio control. This extra mismatch matters because the critic supplies the reference level for every thought token. The advantage asks whether a continuation exceeds its expected remaining return, not merely whether its immediate reward is positive. In a long thought, a small local gain may lead to a promising prefix, while a large gain may leave little useful continuation. GAE then carries later value errors back to earlier tokens (Schulman et al., 2016), so a stale reference can distort credit assignment throughout the thought (Appendix B.2). Before using critic-based advantages, we therefore test both critics on unseen returns from the current rollout law.

4.2 Training and Qualifying Both Critics

To control the critic-specific mismatch above, T^5 trains and qualifies both critics in three steps: stabilize the current return target, test each critic on unseen returns, and repeat the test as the actor changes.

Preparing a stable return target. Within each qualification window, T^5 freezes the actor and its rollout copy, scorer, and reward scale so that two separately parameterized critics $V_i = V_{\psi_i}$, $i \in \{1, 2\}$, face a stable prediction target. We warm up their value heads before adapting the full critics and reset qualification when the backbones are unfrozen. Prediction quality, rather than a fixed warmup duration, determines readiness.

Testing both critics. Let $\mathcal{D}_{\text{qual}} = \{(s_j, G_j)\}_{j=1}^{N_h}$ contain N_h held-out prefix–return observations with nonzero return variance. Each critic receives the predictive score

$$R_i^2 = 1 - \frac{\sum_j (G_j - V_i(s_j))^2}{\sum_j (G_j - \bar{G})^2}, \tag{5}$$

where $\bar{G}$ is the holdout mean. Actor updates require $\min(R_1^2, R_2^2) \geq \eta$, with $0 < \eta < 1$: both critics must reduce the constant predictor's error by at least a fraction η. A score of zero matches that predictor, one fits all held-out returns, and a negative score is worse. Rescaling returns and predictions together leaves the score unchanged.

To see why useful prediction need not have zero loss, write $G - V_i = (G - V^q) + (V^q - V_i)$ at a fixed prefix and scoring item. Since $\mathbb{E}_q[G - V^q \mid s] = 0$, the cross term vanishes:

$$\mathbb{E}_q[(G - V_i(s))^2 \mid s] = \mathrm{Var}_q(G \mid s) + (V_i(s) - V^q(s))^2, \tag{6}$$

where the first term is continuation variability and the second is error in its conditional mean. Qualification tests predictive improvement despite that variability. Taking the weaker score prevents a strong critic from hiding an uninformative partner (Appendix B.2).

Rechecking as the actor changes. Qualification requires $m \geq 1$ consecutive fresh tests on data excluded from fitting. Checks continue at fixed intervals because expected returns after familiar prefixes can change as the actor learns. Sustained failure pauses the actor and resumes critic fitting until qualification is restored. Pausing stabilizes the target distribution, not individual sampled returns, while consecutive passes check that improvement persists across batches. Changes to rollout or scoring targets trigger new checks. Appendix E gives the data separation and update schedule.

The historical naive PPO run illustrates the motivation for a prediction-based warmup. After 20 critic-only steps, actor updates improved CommonsenseQA thought gain from about -4.8 to -3.0 nats at step 40, and mixed gain from about -0.18 to -0.10 nats. Both remained below the no-thought reference (Figure 1). This run used neither T^5 nor the R^2 rule. It provides a diagnostic of fixed-duration warmup, not an evaluation of the proposed method (Appendix A.7).

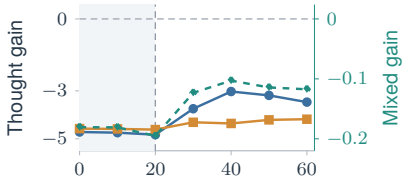


Figure 1: Naive PPO gains (nats) over training steps. Left: CSQA (blue) and GSM8K (orange) thought gains. Right: CSQA mixed gain (teal dashed). Shading: 20-step critic warmup with the actor frozen.

Qualification answers whether each critic predicts current returns. It does not determine how two qualified advantage estimates should be combined, which is the separate problem addressed in Section 5.

5 Learning Complementary Advantages from Single Rollouts

Even after both critics qualify, the actual rollout law q can differ from the recorded policy π_{old} used in the PPO ratio. This training–inference mismatch, as well as clipping, can keep a common advantage offset from canceling in the policy update. We combine the critics' advantages to reduce the offset's contribution while retaining an aggregate learning signal.

5.1 Action-Dependent Mixing and Conditional Drift

With qualified critics in place, we examine how advantages enter clipped PPO, why a common offset can remain, and how the two critics provide room for calibration.

How advantages enter clipped PPO. T^5 retains $\mathcal{J}_{\text{clip}}(\theta)$ from Eq. (4) and changes only its advantage signal. With normalized advantage $\widetilde{A}$ detached, the per-token derivative away from the clipping thresholds is

$$\begin{aligned} \nabla_\theta \min\{\rho\widetilde{A}, \mathrm{clip}(\rho, \rho_-, \rho_+)\widetilde{A}\} &= \widetilde{A}\,\rho I \nabla_\theta \log \pi_\theta(a \mid s) = \widetilde{A} v, \\ I &= \mathbf{1}\{\widetilde{A} > 0, \rho < \rho_+\} + \mathbf{1}\{\widetilde{A} < 0, \rho > \rho_-\} + \mathbf{1}\{\widetilde{A} = 0\}, \end{aligned} \tag{7}$$

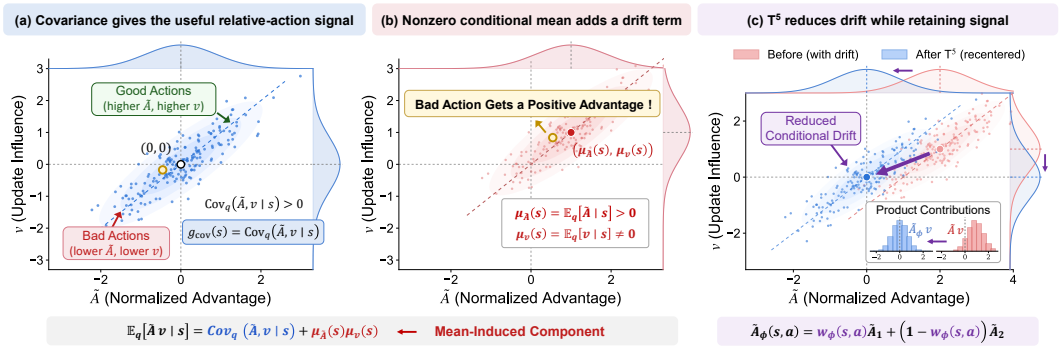


Figure 2: Why calibrate twin-critic advantages? Schematic conditional distributions at a fixed prefix, showing normalized advantage against a scalar projection of the masked update factor v. (a) Centered advantages retain an action-contrast signal. (b) A common offset adds a mean-induced component when the mean update factor is nonzero. (c) Twin-advantage mixing targets the offset while retaining aggregate signal.

where $\mathbf{1}\{\cdot\}$ is the indicator and $v = \rho I \nabla_\theta \log \pi_\theta(a \mid s)$ is the masked update factor. Averaging over replay gives $\nabla_\theta \mathcal{J}_{\text{clip}}(\theta) = \mathbb{E}_{\mathcal{D}}[\widetilde{A} v]$. The advantage supplies sign and strength, while I turns off positive-advantage updates above ρ_+ and negative-advantage updates below ρ_-. Thus PPO stops rewarding further movement once the probability has moved far enough in the favored direction.

Why a common offset matters. Separating each factor into its mean and variation reveals two components of the conditional update:

$$\begin{aligned}\mathbb{E}_q[\widetilde{A} v \mid s] &= \mathbb{E}_q\Big[(\widetilde{A} - \mathbb{E}_q[\widetilde{A} \mid s])(v - \mathbb{E}_q[v \mid s]) \mid s\Big] + \mathbb{E}_q[\widetilde{A} \mid s]\,\mathbb{E}_q[v \mid s] \\ &= \mathrm{Cov}_q(\widetilde{A}, v \mid s) + \mu_{\widetilde{A}}(s)\,\mu_v(s) = \mathrm{Cov}_q(\widetilde{A}, v \mid s) + \Delta_\mu(s),\end{aligned} \tag{8}$$

where $\mu_{\widetilde{A}}(s) = \mathbb{E}_q[\widetilde{A} \mid s]$ is the conditional advantage mean and $\mu_v(s) = \mathbb{E}_q[v \mid s]$ is the conditional mean update factor. Figure 2(a)–(b) contrasts the action-difference signal with the mean-induced component. Under matched sampling without clipping, the score identity gives $\mu_v(s) = 0$. Mismatch or clipping can prevent that cancellation, allowing a common offset to contribute even when action rankings are unchanged. T^5 targets this advantage offset and applies clipping after mixing.

How twin critics enable calibration. Both critics evaluate the same trajectory. Their raw GAE estimates A_i use a common normalization, $\widetilde{A}_i = (A_i - \mu_{\text{pilot}})/\sigma_{\text{pilot}}$, with a shared mean and positive standard deviation fixed from independent pilot data (Appendix E.3). The mixed signal is

$$\widetilde{A}_\varphi = w_\varphi(s, a)\widetilde{A}_1 + (1 - w_\varphi(s, a))\widetilde{A}_2, \tag{9}$$

where $w_\varphi(s, a) \in [0, 1]$ selects between the two estimates using the current action and frozen visible-prefix and critic features, not realized returns or GAEs. The corresponding mask, update factor, and mean-induced component are denoted I_φ, v_φ, and $\Delta_{\mu,\varphi}$.

Figure 2(a)–(b) is a distribution-level schematic, not a tracked-action counterexample. In the concrete construction of Appendix B.4, the clipped-PPO gradient is -0.105 while the true-return gradient is 0.21, showing that the mean-induced component can overwhelm a positive covariance signal. Figure 2(c) illustrates the repair: remove a common offset without erasing the signal. The following two-action example is a separate numerical construction. For two equally likely actions, a fixed rule w can change the twin average $\overline{A}$ into

$$\overline{A} = (0.6, -0.4) \quad \longrightarrow \quad \widetilde{A}_w = (0.6, -0.6), \tag{10}$$

where the mean falls from 0.1 to zero while the action contrast stays nonzero. The signs are unchanged, so the PPO mask and $\mu_v(s)$ remain the same at fixed policy ratios. The mean-induced term therefore vanishes through advantage centering alone. Appendix C.1 gives the original estimates and weights. Section 5.2 specifies an objective for selecting the mixture from single rollouts.

5.2 A SINGLE-ROLLOUT SADDLE-POINT OBJECTIVE

To select the mixture illustrated above from single rollouts, we answer three questions: what objective do we use, why does it work, and what does optimal mixing mean?

What objective do we use? We seek a small squared conditional offset $R(\varphi) = \mathbb{E}_d[\mu_{\widetilde{A}_\varphi}(s)^2]$, not a small advantage at every token. Directly minimizing advantage squares would also suppress within-context variation, including action differences useful to PPO (Appendix C.2). Alongside $w_\varphi(s,a)$, introduce an auxiliary mean function $h(c)$ from a class $\mathcal{H}$. It may use the full pre-action context, including scoring text, but not the sampled action or continuation. With the critics and normalization fixed, we consider the constrained saddle objective

$$\boxed{\min_{\varphi} \sup_{h\in\mathcal{H}} \underbrace{\mathbb{E}_{d,q}[2h(c)\widetilde{A}_\varphi - h(c)^2]}_{\mathcal{L}(\varphi,h)} \quad \text{subject to} \quad C(\varphi) \le \kappa Q}, \tag{11}$$

where $C(\varphi) = \mathbb{E}_{d,q}[(\widetilde{A}_\varphi - \overline{A})^2]$ measures displacement from the fixed average $\overline{A} = (\widetilde{A}_1 + \widetilde{A}_2)/2$, $Q = \mathbb{E}_{d,q}[\overline{A}^2]$ is its squared size, and $0 \le \kappa < 1$ limits their ratio. The mixture minimizes this objective, while the auxiliary function maximizes it. We write h_ζ for a selected auxiliary response, with parameters ζ. Both advantages and the objective's per-token quantity come from one trajectory per selected position, without repeated prefix continuations. Parameterization and optimization details are given in Appendix A.1.

Why does this objective work? For a fixed mixture, completing the square identifies the best mean response. Over all functions with finite $\mathbb{E}_d[h^2]$, denoted $L^2(d)$,

$$R(\varphi) = \sup_{h\in L^2(d)} \mathcal{L}(\varphi,h), \qquad h^*_\varphi(c) = \mu_{\widetilde{A}_\varphi}(s). \tag{12}$$

The maximizing mean function exposes the remaining offset, while the minimizing weight changes the mixture to reduce it. Minimizing over both arguments would instead reward an inaccurate mean response. For a restricted class $\mathcal{H}$, the unresolved mean error enters the bound in Section 5.3. The full square-completion argument is in Appendix C.2. The constraint protects the signal during this correction. The fixed average $w = 1/2$ is feasible, and for $Q > 0$ every feasible mixture satisfies

$$\|\widetilde{A}_\varphi\|_{L^2} \ge (1 - \sqrt{\kappa})\|\overline{A}\|_{L^2} > 0, \tag{13}$$

where $\|X\|_{L^2} = (\mathbb{E}_{d,q}[X^2])^{1/2}$. Feasible mixtures lie in a ball around the fixed average that excludes zero. Thus a smaller conditional mean need not come from erasing the aggregate learning signal. The reverse triangle inequality proves this in one step (Appendix C.5).

What does optimal mixing mean? Let $x = (s,a)$ denote only the gate's visible inputs and $\Delta\widetilde{A} = \widetilde{A}_1 - \widetilde{A}_2$. The population optimum has the following form.

Proposition 1 (Optimal mixing)**.** *Fix a finite nonzero rollout occupation measure and square-integrable advantages, with $Q > 0$ and $0 < \kappa < 1$. Over all measurable gates $w(x) \in [0,1]$ and $h \in L^2(d)$, the constrained problem has a Lagrangian saddle point (w^*, h^*, β^*), where $h^*(c) = \mathbb{E}_q[\widetilde{A}_{w^*} \mid c]$ and $\beta^* \ge 0$ is the retention multiplier. Wherever $\beta^* > 0$ and $\mathbb{E}_{d,q}[(\Delta\widetilde{A})^2 \mid x] > 0$,*

$$w^*(x) = \mathrm{clip}\left(\frac{1}{2} - \frac{\mathbb{E}_{d,q}[h^*(c)\Delta\widetilde{A} \mid x]}{\beta^*\mathbb{E}_{d,q}[(\Delta\widetilde{A})^2 \mid x]},\, 0, 1\right). \tag{14}$$

The numerator sets the adjustment direction through the relation between offset and critic disagreement. The denominator scales it by disagreement and retention pressure, and projection keeps the result between the two estimates. This self-consistent relation characterizes the population optimum, not an extra prefix-wise estimation step. Strong duality, the precise function-space conditions, and boundary cases are given in Appendix C.4.

Algorithm 1 places this objective within PPO. The selected mixture remains fixed during actor updates, and the complete mixed advantage is detached. Appendix A.1 describes the parameterized realization and its fitting and validation procedure.

Algorithm 1 T^5: one training window

Require: Actor π_θ, critics V_1, V_2, displacement limit κ.
1: Warm up the value heads, then the full critics using $\mathcal{L}_{V_i}$ (Eq. (118)). *// Critic fitting*
2: Require m fresh tests with $\min(R_1^2, R_2^2) \geq \eta$ (Eq. (5)). If not, pause the actor and refit.
3: Learn w_φ, h_ζ (Eq. (11)). *// Advantage calibration*
4: Validate and freeze w_φ (Appendix C.7).
5: Sample fresh trajectories and form $\widetilde{A}_\varphi$ (Eq. (9)). *// Policy optimization*
6: Update θ (Eq. (4)).

Each selected position supplies one thought to both critics. Additional critic-fitting and mixture-validation trajectories are included in the group-sampling cost comparison in Appendix E.7.

5.3 Bounding Conditional Drift

For fitted heads, two quantities connect the learned objective to residual drift. The conditional-mean prediction error is $\epsilon_h := \mathbb{E}_d[(h_\zeta(c) - \mu_{\widetilde{A}_\varphi}(s))^2]$. An independent validation estimate $\widehat{\mathcal{L}}_{\mathrm{val}}$ measures the objective, with uncertainty ϵ_{stat}. The first quantity concerns the true conditional mean, not regression error against individual sampled advantages. Its approximation and optimization components are detailed in Appendix C.2.

Theorem 1. *Condition on the frozen rollout law, critics, scorer, normalizer, and selected head parameters. Assume square-integrable advantages and auxiliary functions, valid recorded probabilities, differentiable policies, and the required finite score moments. Let B bound the weighted size of the mean update factor, $\mathbb{E}_d\|\mu_{v_\varphi}\|_2^2 \leq B^2$, where $\|\cdot\|_2$ is the Euclidean norm. For a frozen candidate and an independent validation estimate with $\mathcal{L}(\varphi, h_\zeta) \leq \widehat{\mathcal{L}}_{\mathrm{val}} + \epsilon_{\mathrm{stat}}$,*

$$\mathbb{E}_d\|\Delta_{\mu,\varphi}\|_2 \leq B\sqrt{R(\varphi)} \leq B\sqrt{\max\{0, \widehat{\mathcal{L}}_{\mathrm{val}} + \epsilon_{\mathrm{stat}} + \epsilon_h\}}. \tag{15}$$

The bound links the remaining advantage offset to its amplification by the policy update. Cauchy–Schwarz gives the first inequality, and the dual identity with validation gives the second (Appendix C.6). A small measured objective gives a tight bound when the mean-prediction error and validation uncertainty are also small. This controls the mean-induced component with the actual PPO mask, rather than the entire update. Appendix C.7 specifies the validation conditions.

6 Experiments

We organize the experiments around three questions:

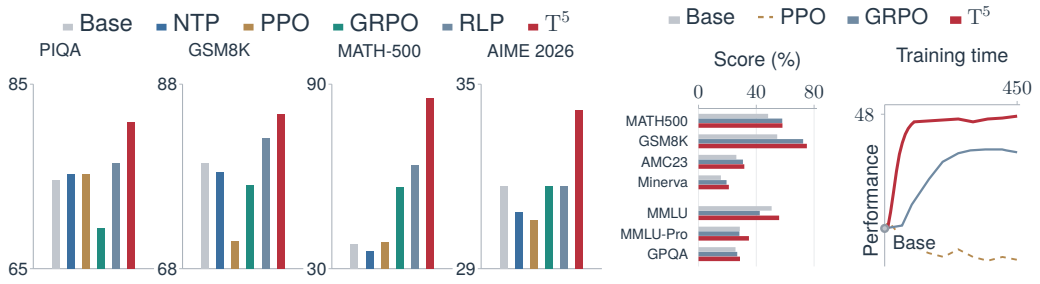
1. Does T^5 improve downstream performance over representative strong baselines?
2. Are the gains consistent across tasks and evaluation settings?
3. How does end-to-end training efficiency scale with model size?

6.1 Experimental Setup

We continue pretraining Qwen3.5-2B-Base and Qwen3.5-9B-Base on a fixed corpus with disjoint data splits (Qwen Team, 2026). The 2B study compares Base, NTP, single-critic Quiet-STaR (w/ PPO), Quiet-STaR (w/ GRPO), RLP, and T^5 under matched settings (Zelikman et al., 2024; Hatamizadeh et al., 2026), while 9B measures scaling. All GRPO-based baselines use group size 8. Appendix A gives the full protocol, timing traces, and compute accounting.

6.2 Main Results

Overall downstream effectiveness. Table 7 shows that T^5 leads all seven retained benchmarks. Figure 3(a) highlights four with the clearest separation. Its unweighted mean is 69.56, versus 64.52 for RLP and 60.15 for Quiet-STaR (w/ GRPO). The largest RLP margin is on MATH-500 (85.45 versus 63.65). Gains on the other six tasks are 0.90–4.51 percentage points.



(a) Benchmark results **(b)** Controlled comparison

Figure 3: Experimental comparisons. (a) Selected Qwen3.5-2B results (PIQA: accuracy; others: pass@1). (b) Earlier controlled comparison with a shared Quiet-STaR architecture: benchmark scores (%, left) and performance over training time from the common Base checkpoint (right). Legend labels PPO and GRPO denote Quiet-STaR (w/ PPO) and Quiet-STaR (w/ GRPO), respectively.

After correcting its AIME 2025/2026 scores to 33.00/30.83, NTP averages 56.30 and is nearly unchanged from Base (56.28). Quiet-STaR (w/ PPO) averages 54.42, below Base in this run. Without seed-level uncertainty, these point estimates establish only the observed ordering, not statistical significance or component-level causality. Appendix A.3 gives the rounded table, Pass@4, reasoning-token diagnostics, and provenance.

Cross-task consistency. Before the full target protocol, we use an earlier controlled comparison to ask whether the advantage persists across mathematical and knowledge-intensive tasks (Yang et al., 2025). All trained variants share the Quiet-STaR hidden-thought architecture. The two policy-optimization baselines are therefore denoted Quiet-STaR (w/ PPO) and Quiet-STaR (w/ GRPO) (Zelikman et al., 2024; Shao et al., 2024).

Figure 3(b) compares Base, Quiet-STaR (w/ GRPO), and T^5 on seven mathematical and knowledge-intensive tasks. T^5 exceeds Quiet-STaR (w/ GRPO) on all seven, with a small difference on MATH500 and larger differences on GSM8K and MMLU. It also exceeds Base on the displayed tasks, while Quiet-STaR (w/ GRPO) improves mathematics but falls 8.21 points below Base on MMLU. The full summary includes an exception: T^5 is slightly below Base on GPQA pass@1. Appendix A.4 gives task references, all scores, and source information. These point estimates motivate controlled evaluation but do not isolate predictive qualification or complementary weighting.

Training-time trade-off. Group sampling, value fitting, and calibration incur different costs. Against RLP, Table 1 shows mean step-time reductions of 41.1% at 2B and 63.4% at 9B, with cross-scale performance in Appendix A.6. Separately, Figure 3(b) traces the earlier Quiet-STaR (w/ PPO), Quiet-STaR (w/ GRPO), and T^5 runs from their common Base checkpoint. Timings use the same hardware and include critic warmup, thought generation, reward scoring, critic fitting, weight and mean learning, validation, and rejected windows. The relevant question is whether maintaining useful actor signals compensates for the cost of two critics and auxiliary continuations. Appendix A gives the full reporting protocol.

Table 1: Mean step time (s).

Model	RLP	T^5
Qwen3.5-2B	54.01	31.79
Qwen3.5-9B	497.83	182.18

7 Conclusion

T^5 separates return prediction from advantage calibration. Qualified twin critics supply estimates that a conditional-moment saddle objective combines using one rollout per selected position. Its population optimality conditions explain how the remaining offset, critic disagreement, and retention constraint determine the mixture. The retention constraint preserves aggregate signal strength, and the analysis bounds the mean-induced component of clipped PPO drift. This makes both return prediction and the use of its advantage estimates part of the learning process.

Reproducibility Statement

The method and training objectives are specified in the main text, with derivations in Appendices B.2 and C. Appendix E details data separation, parameter updates, and trajectory counts. Appendix A documents the evaluation protocol and the provenance of the preliminary scores and historical diagnostic. The coordinates underlying the training-time figure are retained with the plotting data.

References

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. PIQA: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020. URL https://arxiv.org/abs/1911.11641.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018. URL https://arxiv.org/abs/1803.05457.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Qingxiu Dong, Li Dong, Yao Tang, Tianzhu Ye, Yutao Sun, Zhifang Sui, and Furu Wei. Reinforcement pre-training. *arXiv preprint arXiv:2506.08007*, 2025.

Scott Fujimoto, Herke van Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 1587–1596, 2018.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 1861–1870, 2018.

Ali Hatamizadeh, Syeda Nahida Akter, Shrimai Prabhumoye, Jan Kautz, Mostofa Patwary, Mohammad Shoeybi, Bryan Catanzaro, and Yejin Choi. RLP: Reinforcement as a pretraining objective. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2510.01265.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021a.

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. *NeurIPS*, 2021b.

Zhenyu Hou, Yujiang Li, Jie Tang, and Yuxiao Dong. Single-rollout asynchronous optimization for agentic reinforcement learning. *arXiv preprint arXiv:2607.07508*, 2026. URL https://arxiv.org/abs/2607.07508.

Jian Hu, Jason Klein Liu, Haotian Xu, and Wei Shen. REINFORCE++: Stabilizing critic-free policy optimization with global advantage normalization. *arXiv preprint arXiv:2501.03262*, 2025.

Wei Huang, Yizhe Xiong, Xin Ye, Zhijie Deng, Hui Chen, Zijia Lin, and Guiguang Ding. Fast quiet-STar: Thinking without thought tokens. *arXiv preprint arXiv:2505.17746*, 2025.

Vijay R. Konda and John N. Tsitsiklis. On actor-critic algorithms. *SIAM Journal on Control and Optimization*, 42(4):1143–1166, 2003. doi: 10.1137/S0363012901385691.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. Solving quantitative reasoning problems with language models. *Advances in Neural Information Processing Systems*, 35:3843–3857, 2022.

Siheng Li, Kejiao Li, Zenan Xu, Guanhua Huang, Evander Yang, Kun Li, Haoyuan Wu, Jiajia Wu, Zihao Zheng, Chenchen Zhang, et al. Reinforcement learning on pre-training data. *arXiv preprint arXiv:2509.19249*, 2025.

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. *arXiv preprint arXiv:2503.20783*, 2025.

Renjie Mao, Xiangxin Zhou, Lvfang Tao, Yixin Ding, Yu Shi, Yongguang Lin, Yuheng Wu, Honglin Zhu, Qian Qiu, and Wenxi Zhu. Beyond uniform token-level trust region in LLM reinforcement learning. *arXiv preprint arXiv:2606.10968*, 2026.

Martin Marek and Max Ryabinin. Score centering stabilizes off-policy reinforcement learning. *arXiv preprint arXiv:2609.20807*, 2026.

Mathematical Association of America. American invitational mathematics examination (AIME). Mathematics Competition Series, n.d. URL https://maa.org/math-competitions/aime.

Andrew Y. Ng, Daishi Harada, and Stuart J. Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the 16th International Conference on Machine Learning*, pp. 278–287, 1999.

Keiran Paster, Marco Dos Santos, Zhangir Azerbayev, and Jimmy Ba. OpenWebMath: An open dataset of high-quality mathematical web text. *arXiv preprint arXiv:2310.06786*, 2023.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro von Werra, and Thomas Wolf. The FineWeb datasets: Decanting the web for the finest text data at scale. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://arxiv.org/abs/2406.17557.

Nan Qiao, Sheng Yue, Shuning Wang, Yongheng Deng, and Ju Ren. Less is more: Clustered cross-covariance control for offline RL. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://arxiv.org/abs/2601.20765.

Nan Qiao, Sheng Yue, Shuning Wang, and Ju Ren. AdamO: A collapse-suppressed optimizer for offline RL. In *International Conference on Machine Learning*, 2026b. URL https://arxiv.org/abs/2605.01968.

Qwen Team. Qwen3.5: Towards native multimodal agents. Qwen Blog, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 1889–1897, 2015.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1506.02438.

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

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

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, pp. 1057–1063, 2000.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pp. 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1421. URL https://aclanthology.org/N19-1421.

Yijun Tian, Shaoyu Chen, Zhichao Xu, Yawei Wang, Jinhe Bi, Peng Han, and Wei Wang. Reinforcement mid-training. *arXiv preprint arXiv:2509.24375*, 2025.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290, 2024.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3–4):229–256, 1992.

Xingrun Xing, Zhiyuan Fan, Jie Lou, Guoqi Li, Jiajun Zhang, and Debing Zhang. Pretrainzero: Reinforcement active pretraining. *arXiv preprint arXiv:2512.03442*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, et al. DAPO: An open-source LLM reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Yu Yue, Yufeng Yuan, Qiying Yu, Xiaochen Zuo, Ruofei Zhu, Wenyuan Xu, Jiaze Chen, Chengyi Wang, Tiantian Fan, Zhengyin Du, et al. VAPO: Efficient and reliable reinforcement learning for advanced reasoning tasks. *arXiv preprint arXiv:2504.05118*, 2025.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. STaR: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 15476–15488, 2022.

Eric Zelikman, Georges Raif Harik, Yijia Shao, Varuna Jayasiri, Nick Haber, and Noah Goodman. Quiet-STar: Language models can teach themselves to think before speaking. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=oRXPiSOGH9.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

FULL PAPER CONTENTS

1 Introduction **1**

2 Related Work **2**

3 Preliminaries and Problem Setup **3**

3.1 Reinforcement Mid-Training from Unlabeled Text 3

3.2 PPO with a Learned Token-Level Critic 3

4 Critic Mismatch and Predictive Qualification **4**

4.1 The Critic Adds Its Own Mismatch 4

4.2 Training and Qualifying Both Critics 4

5 Learning Complementary Advantages from Single Rollouts **5**

5.1 Action-Dependent Mixing and Conditional Drift 5

5.2 A Single-Rollout Saddle-Point Objective 7

5.3 Bounding Conditional Drift 8

6 Experiments **8**

6.1 Experimental Setup 8

6.2 Main Results 8

7 Conclusion **9**

A Additional Experimental Details and Results **15**

A.1 Model, Data, and Implementation Configuration 15

A.2 Target Evaluation and Comparison Protocol 16

A.3 Qwen3.5-2B Results and Diagnostics 17

A.4 Earlier Cross-Task Comparison and Provenance 20

A.5 Recorded Per-Step Training Time 21

A.6 Performance and Efficiency Across Model Scales 22

A.7 Historical Naive-PPO Diagnostic 24

B Why Predictive Accuracy Does Not Remove Drift **25**

B.1 What changes when old rollouts train a new policy? 25

B.2 Value error and predictive qualification 28

B.3 Two distinct consistency conditions for cancellation 30

B.4 Why a fixed minimum does not calibrate advantages 31

C Conditional-Moment Learning and the Drift Bound **33**

C.1 Action dependence and the limits of twin disagreement 33

C.2 The conditional-moment dual 34
C.3 One-rollout stochastic gradients 34
C.4 Population saddle points and optimal weights 35
C.5 Signal retention and attainable risk 36
C.6 Actual PPO gradient and error bounds 37
C.7 Independent trajectory validation and its limits 38
C.8 Policy sensitivity, clipping, and sampling mismatch 38

D Scope of Conditional Calibration **40**
D.1 Why one-step TD is not substituted for GAE 40
D.2 Training-time scoring information and visible-state limits 40
D.3 Why qualification must follow the prefix distribution 41
D.4 Relation to Neighboring Actor–Critic Methods 42

E Training Protocol and Computational Cost **43**
E.1 Reward scale and checkpoint construction 43
E.2 Value regression and snapshot references 43
E.3 Advantage snapshots and masked GAE 44
E.4 Trajectory collection and actor replay 44
E.5 The prediction-mixture gate 45
E.6 State, action, and computational cost 45
E.7 Trajectory counts and computational cost 45
E.8 Refreshing and recovering a training window 46

A Additional Experimental Details and Results

This appendix first gives the model, data, and evaluation protocols, then follows the three questions in Section 6: overall performance, cross-task consistency, and training efficiency across model scales. Historical training diagnostics follow these results.

A.1 Model, Data, and Implementation Configuration

The target protocol uses the two 2026 Qwen3.5 Base checkpoints in Table 2 (Qwen Team, 2026).[1] Approximate file sizes describe checkpoint storage, not training memory. Neither initialization is an Instruct checkpoint.

Table 2: Target Base-model initializations.

Checkpoint	Parameters	Storage
`Qwen/Qwen3.5-2B-Base`	2.274B	4.56 GB
`Qwen/Qwen3.5-9B-Base`	9.653B	19.32 GB

The corpus release is `recent_fast_100step_v1`. Table 3 specifies the source mixture, and Table 4 preserves the supplied sequence and corpus-token counts. FineWeb supplies web text (Penedo et al., 2024),[2] arXiv supplies 2025–2026 titles and abstracts, and OpenWebMath supplies mathematical web text (Paster et al., 2023). The supplied configuration calls the scholarly component `open-index/open-arxiv`. Source dates alone cannot rule out benchmark contamination.

Table 3: Training-corpus composition. The four web crawls are from 2025.

Source	Share
FineWeb: `CC-MAIN-2025-13/18/21/26`	85%
2025–2026 open-index/open-arxiv titles and abstracts	10%
OpenWebMath	5%

Table 4: Corpus splits, with 512 tokens per sequence. Counts exclude generated thoughts and auxiliary continuations.

Split	Sequences	Corpus tokens
Training	12,800	6,553,600
Critic calibration	1,024	524,288
Development	1,024	524,288

Actor optimization and critic return fitting use the training split. Predictive qualification uses the calibration split, with held-out outcomes excluded from fitting and from selecting a threshold on those outcomes. Development data support progress measurement and a checkpoint-selection rule fixed before final evaluation. Complete text items or parent trajectories are partitioned into the roles in Appendix E before token batching. Head learning, independent validation, and actor optimization use separate ordinary trajectories, each with one thought per selected item. Predictive qualification alone does not certify the learned mixture's conditional moment risk.

[1] Model resources: `https://huggingface.co/Qwen/Qwen3.5-2B-Base` and `https://huggingface.co/Qwen/Qwen3.5-9B-Base`. These identifiers specify the selected initializations.

[2] FineWeb dataset: `https://huggingface.co/datasets/HuggingFaceFW/fineweb`. The selected crawl identifiers are listed in Table 3.

Table 5: Methods in the reported Qwen3.5-2B comparison. A dash indicates no grouped thought sampling. All GRPO-based comparisons in this paper use group size 8.

Method	Group	Update	Controlled role
Base	–	None	Original checkpoint with no additional corpus exposure.
NTP	–	Next-token likelihood	Controls for continued exposure to the same training mixture.
Quiet-STaR (w/ PPO)	1	Single-critic clipped PPO	Uses the hidden-thought interface with a standard learned-value update.
RLP	8	GRPO-style group-relative update	Samples eight thoughts and uses next-token information-gain rewards.
Quiet-STaR (w/ GRPO)	8	Group-relative update	Applies GRPO to the same hidden-thought architecture.
T^5	1	Twin-critic clipped PPO	Uses predictive qualification and constrained saddle-point calibration.

Execution environment. The target execution environment uses three nodes with NVIDIA H20 GPUs, CUDA 12.9, and Python 3.13.

Calibration parameterization and optimization. The functions in Section 5.2 can be parameterized by lightweight MLP heads rather than additional language-model backbones. The weight head $w_\varphi(s, a)$ has output in $[0, 1]$ and uses frozen visible-prefix and critic features together with the current action. The scalar mean head $h_\zeta(c)$ can additionally use the scoring context through a separate training-only input path, but not the sampled action or its continuation. Returns and GAEs are labels, not inputs. MLP depth, width, activations, feature encodings, optimizer settings, and update ratios are run-specific choices to record for the implementation, rather than requirements of the objective.

During calibration, fix the rollout law, critics, scorer, and common pilot normalizer, and detach the GAE labels and critic features. Minimize in φ and maximize in ζ and the nonnegative retention multiplier, using the single-trajectory gradients in Appendix C.3. Each selected item contributes one thought, shared by both critics. Freeze the candidate before independent validation, then construct detached advantages on fresh actor trajectories. A failed retention check uses the fixed average, and failed critic qualification pauses actor updates. Appendix E.4 specifies the trajectory-level separation and replay rules. This parameterized fitting procedure realizes the saddle objective without changing the actor's clipped PPO loss.

A.2 Target Evaluation and Comparison Protocol

Table 6 defines the seven-benchmark primary suite. AIME 2025 and 2026 use the official competition sets (Mathematical Association of America, n.d.). GSM8K (Cobbe et al., 2021) and MATH-500 (Hendrycks et al., 2021b) cover mathematical reasoning. CommonsenseQA (Talmor et al., 2019), PIQA (Bisk et al., 2020), and ARC-Challenge (Clark et al., 2018) cover general, physical, and scientific commonsense.

Table 6: Primary evaluation suite and frozen item counts.

Release	Benchmark	Items	Focus
2025	AIME 2025	30	Competition mathematics
2026	AIME 2026	30	Competition mathematics
2021	GSM8K test	1,319	Grade-school mathematics
2021	MATH-500	500	Competition mathematics
2019	CommonsenseQA validation	1,221	General commonsense
2020	PIQA validation	1,838	Physical commonsense
2018	ARC-Challenge test	1,172	Science reasoning

AIME, GSM8K, and MATH-500 use fixed answer extraction and normalization before exact-match scoring. CommonsenseQA, PIQA, and ARC-Challenge use accuracy on their frozen multiple-

choice splits. All primary benchmarks are excluded from training, critic fitting, admission, calibration, and checkpoint selection, and score interpretation requires corpus-overlap checks.

Within each model scale, methods share the same Base initialization, corpus, selected positions, evaluation prompts, demonstrations, output-token limit, decoding parameters, and tool or retrieval allowances. Base checkpoints use an explicit prompting protocol rather than an assumed Instruct template. The 512-token training sequence does not determine evaluation generation length. The reported results are point estimates without variation across independent training seeds, which is especially limiting for the 30-item AIME sets.

The comparison roles are summarized in Table 5. NTP controls corpus exposure, while Quiet-STaR (w/ PPO) and Quiet-STaR (w/ GRPO) compare two policy-update choices on the hidden-thought architecture. Group-8 RLP provides the closest information-gain baseline (Hatamizadeh et al., 2026). Comparisons distinguish matched corpus tokens, sampled thought tokens, and total compute. Compute accounting follows Eq. (123) and includes trajectory-set sizes, generated trajectories and tokens, actor-eligible tokens, group sizes, and head-update counts. Head-training and validation trajectories count toward generation without contributing actor tokens. Total cost includes selected positions, thought length, candidate generation, reward scoring, both critics, head-learning and validation sets, rejected windows, and hyperparameter search. Equal optimizer steps do not imply equal compute.

Development diagnostics include continuation loss, thought and mixed NLL gain, useful-position fraction, and worst-1% gain. Each critic is evaluated by held-out R^2, prediction standard deviation, and twin disagreement. Policy KL, clipping frequency, actor-release timing, the restricted moment objective, signal-retention ratio, and head/validation cost characterize the new protocol. Reproducibility additionally requires optimizer settings, batch and accumulation sizes, precision, hardware, trainable parameters, seeds, thought-length limits, head architectures and inputs, κ, and split rules.

A.3 Qwen3.5-2B Results and Diagnostics

Primary benchmark results. The supplied Qwen3.5-2B result matrix contains the six methods in Table 5: Base, NTP, Quiet-STaR (w/ PPO), RLP, Quiet-STaR (w/ GRPO), and T^5. Table 7 reports accuracy for CommonsenseQA, PIQA, and ARC-Challenge and pass@1 for the four mathematical tasks. The values are author-supplied point estimates.

Table 7: Qwen3.5-2B primary results on the seven retained benchmarks, rounded to integer percentages. Bold and underlined entries identify the best and second-best unrounded values in each row. The final row rounds the unweighted mean of the seven source values.

Benchmark	Base	NTP	Quiet-STaR (w/ PPO)	RLP	Quiet-STaR (w/ GRPO)	T^5
CSQA	66	67	64	<u>74</u>	67	**75**
PIQA	75	75	75	<u>76</u>	69	**81**
ARC-Challenge	74	74	75	<u>88</u>	86	**90**
GSM8K	79	78	71	<u>82</u>	77	**85**
MATH-500	38	36	39	<u>64</u>	56	**85**
AIME 2025	30	33	27	<u>36</u>	33	**38**
AIME 2026	32	31	31	<u>32</u>	<u>32</u>	**34**
Retained-7 mean	56	56	54	<u>65</u>	60	**70**

The current means are 56.28 for Base, 56.30 for NTP, 54.42 for Quiet-STaR (w/ PPO), 64.52 for RLP, 60.15 for Quiet-STaR (w/ GRPO), and 69.56 for T^5. T^5 leads on all seven tasks and exceeds RLP by 5.04 points on the retained-seven mean. RLP exceeds Quiet-STaR (w/ GRPO) on six tasks and ties it on AIME 2026. Figure 3(a) in the main text displays the unrounded values for four selected tasks.

Aggregation note. The retained-seven mean is recomputed from the unrounded source values in Table 7. Earlier aggregate values of 55.95 for RLP and 51.57 for Quiet-STaR (w/ GRPO) covered

a broader suite that included benchmarks later removed as obsolete, so they are not mixed with the current seven-task summary. The NTP scores on AIME 2025/2026 are corrected to 33.00/30.83.

Sampling gains and reasoning length. Table 8 records the additional T^5 diagnostics. Coverage is 100% on every benchmark. Pass@4 is available for the four generative mathematics evaluations and exceeds pass@1 by 15.83 points on each AIME set, 8.38 points on GSM8K, and 4.15 points on MATH-500. The average reasoning length varies substantially by task, from 548.2 tokens on PIQA to 3922.4 on AIME 2026. Because these token measurements are available only for T^5, they describe its cross-task inference profile rather than a cross-method efficiency comparison.

Table 8: Qwen3.5-2B T^5 evaluation diagnostics. Primary scores, pass@4, and coverage are percentages. A dash marks tasks for which pass@4 was not supplied.

Benchmark	Metric	Primary	Pass@4	Coverage	Avg. tokens
AIME 2025	pass@1	37.50	53.33	100.00	3849.1
AIME 2026	pass@1	34.17	50.00	100.00	3922.4
ARC-Challenge	Accuracy	89.59	–	100.00	609.2
CSQA	Accuracy	74.53	–	100.00	649.2
GSM8K	pass@1	84.80	93.18	100.00	645.7
MATH-500	pass@1	85.45	89.60	100.00	1673.6
PIQA	Accuracy	80.90	–	100.00	548.2

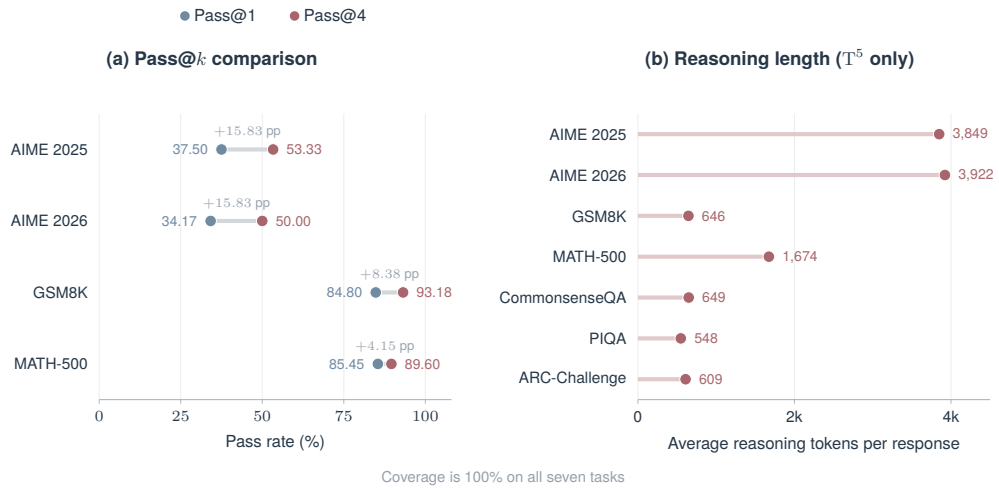


Figure 4: Qwen3.5-2B T^5 diagnostics. (a) Paired pass@1 and pass@4 with percentage-point differences. (b) Average reasoning tokens per response, with rounded endpoint labels. Coverage is 100% on all seven tasks.

Warmup parameter ablation. We vary the warmup parameter, denoted by τ in this sweep, over $\{0.00, 0.05, 0.10, 0.20, 0.40\}$ on Qwen3.5-2B and evaluate AIME 2026. Figure 5 pairs each score with its measured warmup time. The score rises from 29.9 to 35.0 as warmup increases from 1.2 to 23.1 minutes. Gains diminish at the upper end: moving from $\tau = 0.20$ to 0.40 adds only 0.2 percentage points while requiring 12.4 additional minutes. The $\tau = 0.00$ setting still incurs 1.2 minutes of warmup, rather than removing warmup entirely.

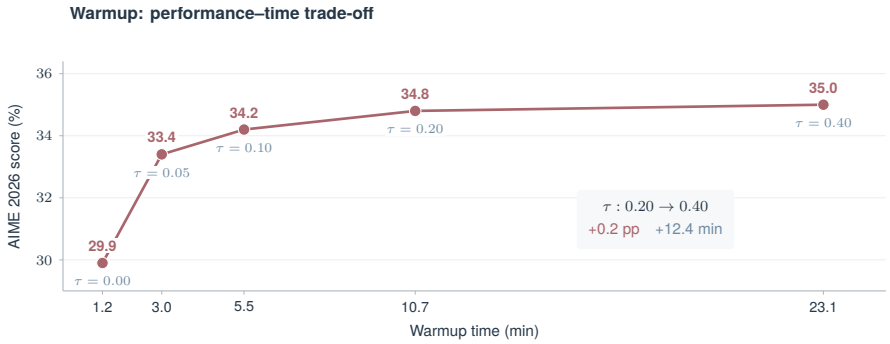


Figure 5: Qwen3.5-2B T^5 warmup-parameter ablation on AIME 2026. Scores are plotted against measured warmup time, with point labels identifying the score and τ.

A.4 Earlier Cross-Task Comparison and Provenance

The available summary compares Base, Quiet-STaR (w/ GRPO), and T^5 on Qwen3-1.7B-Base (Yang et al., 2025). All trained variants use the same Quiet-STaR hidden-thought architecture. The group-relative variant uses GRPO (Zelikman et al., 2024; Shao et al., 2024). The mathematics group comprises MATH500, drawn from MATH (Hendrycks et al., 2021b), GSM8K (Cobbe et al., 2021), AMC23, and Minerva (Lewkowycz et al., 2022). The other tasks are MMLU (Hendrycks et al., 2021a), MMLU-Pro (Wang et al., 2024), and GPQA (Rein et al., 2024). This suite differs from the primary target suite in Section 6. Its Quiet-STaR (w/ GRPO) baseline also uses group size 8 for both the benchmark summary and the training-time comparison.

Table 9 preserves all ten benchmark entries in the supplied summary. The scores are numerical inputs to the original plotting script, rather than newly computed evaluations. The Qwen3-1.7B initialization and shared architecture are author-confirmed, while the accompanying machine-readable materials do not record the training corpus, inference settings, training seeds, or exact T^5 implementation. Consequently, these results are not attributed to either Qwen3.5 initialization or to the single-rollout conditional-moment protocol analyzed in this paper.

Table 9: Complete preliminary benchmark summary (%). Bold indicates the largest reported value in each row. The three additional rows report pass@1.

Benchmark	Base	Quiet-STaR (w/ GRPO)	T^5
MATH500	48.17	57.96	**58.10**
GSM8K	54.47	72.42	**75.00**
AMC23	26.19	30.76	**31.70**
Minerva	15.50	19.40	**21.00**
MMLU	50.61	42.40	**55.80**
MMLU pass@1	45.10	40.49	**51.90**
MMLU-Pro	28.60	28.30	**34.90**
MMLU-Pro pass@1	24.30	25.10	**31.10**
GPQA	25.60	26.80	**28.70**
GPQA pass@1	**27.00**	25.20	26.80

Figure 3(b) shows higher T^5 scores than Quiet-STaR (w/ GRPO) on all seven displayed tasks. The difference is small on MATH500 (58.10 versus 57.96), but larger on GSM8K (75.00 versus 72.42) and MMLU (55.80 versus 42.40). Relative to Base, Quiet-STaR (w/ GRPO) improves the four mathematics scores while its MMLU score decreases by 8.21 points, from 50.61 to 42.40. T^5 exceeds Base on the seven displayed tasks. The complete summary also contains an exception: on GPQA pass@1, T^5 scores 26.80 against Base's 27.00. Thus the observed improvement is not uniform across all reported evaluation entries. The Base scores are not uniformly depressed: on the seven main tasks they differ by at most 0.53 points from the Qwen3-1.7B-Base values reported with RLP (Hatamizadeh et al., 2026). The conspicuous value is instead the Quiet-STaR (w/ GRPO) MMLU regression, which we treat as a run-specific observation rather than general evidence.

The three additional rows report pass@1. The supplied summary does not identify their exact aggregation protocol. The benchmark panel shows the seven main entries, while the table retains the three pass@1 values, including the unfavorable GPQA comparison with Base. These point estimates motivate a controlled comparison but do not isolate predictive gating or complementary weighting.

Figure 3(b) plots the supplied Quiet-STaR (w/ PPO), Quiet-STaR (w/ GRPO), and T^5 performance trajectories against training time. Their common time-zero point denotes the shared Qwen3-1.7B-Base checkpoint. The remaining points belong to the corresponding method. The compact panel omits numeric ticks and intermediate markers, while the source JSON retains every coordinate. Timing includes critic warmup, rollout generation, reward scoring, critic fitting, weight and auxiliary-head learning, validation, and rejected windows under matched hardware and evaluation settings.

A.5 Recorded Per-Step Training Time

Figure 6 compares T^5 with the critic-free RLP baseline, which uses GRPO-style updates with group size 8 at both model scales. It shows every supplied `step_time_seconds` value as a faint line, with 15-step rolling medians overlaid. The two 2B RLP exports cover steps 1–200 and 201–500, respectively. We join them at their original step indices and use all 500 records for the reported RLP mean. Table 1 uses the final continuous 2B T^5 segment (steps 301–600) for its mean, excluding the earlier startup and interrupted segments. For T^5, the figure retains all three column pairs in timestamp-defined order. Because the second segment restarts its local step count, later segments are offset to form a monotone plotted step axis. Dashed vertical lines mark the two joins. The logarithmic time axis retains the initial high-cost segment without flattening the remaining measurements.

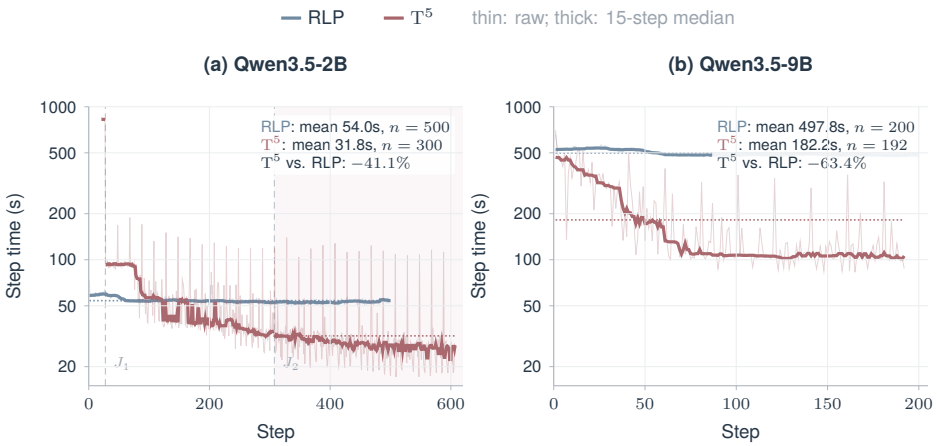


Figure 6: Recorded per-step training times for RLP and T^5 on Qwen3.5-2B and Qwen3.5-9B. Faint lines show raw timers, solid lines show 15-step rolling medians, and dotted lines mark the reported means. The shaded 2B region is the final continuous T^5 segment used for its mean; J_1 and J_2 mark segment joins. The logarithmic vertical axis reports seconds.

A.6 Performance and Efficiency Across Model Scales

Within-scale comparison. Figure 7 compares Base, NTP, and T^5 on Qwen3.5-9B using the seven-task metric convention of Table 7. T^5 exceeds both baselines on every task. Its unweighted mean is 92.71, compared with 86.86 for Base and 87.00 for NTP, giving gains of 5.86 and 5.71 percentage points, respectively. The largest margin over NTP is on AIME 2026 (93 versus 76), while MATH-500 improves from an already high 96 to 97. These results show that the method remains useful when applied to a stronger initialization.

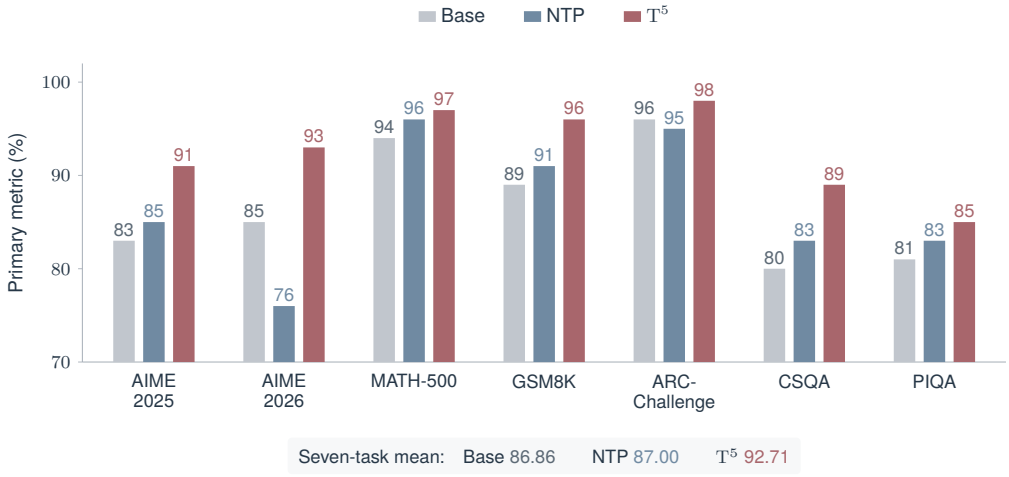


Figure 7: Qwen3.5-9B comparison of Base, NTP, and T^5. Labels preserve the supplied scores in percent. The four mathematics tasks use pass@1, and the three multiple-choice tasks use accuracy. CSQA abbreviates CommonsenseQA. The vertical axis starts at 70, and the footer reports unweighted means over all seven tasks.

Cross-scale performance. Figure 8 aligns the two model scales using only the three shared methods. All three improve on every task at 9B. The T^5 mean rises from 69.56 to 92.71, a 23.15-point increase. Its margin over NTP remains positive at both scales, although it narrows from 13.26 to 5.71 points as the baseline scores rise. The comparison therefore separates the benefit of a larger base model from the additional benefit of T^5 within each scale.

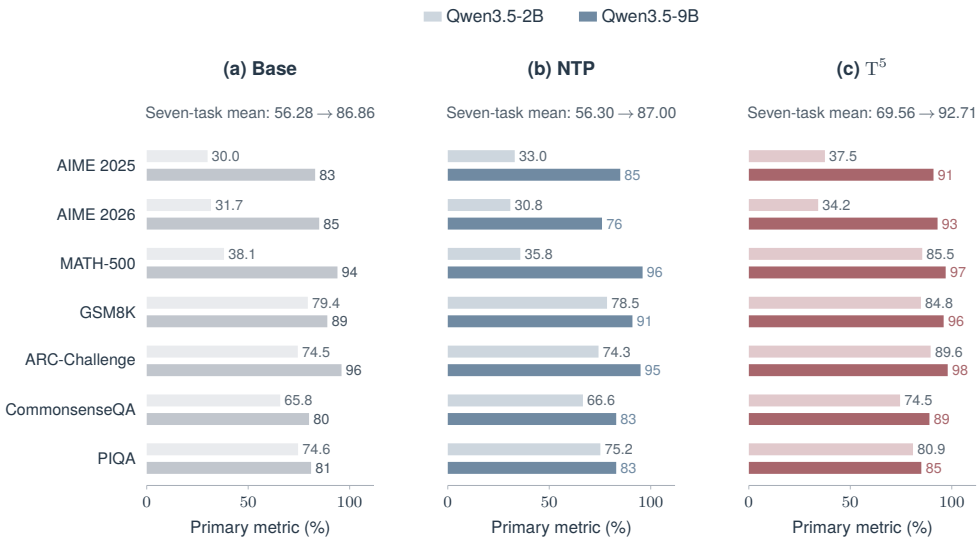


Figure 8: Common-method comparison between Qwen3.5-2B and Qwen3.5-9B. Each panel pairs a method's scores on the same seven tasks. Light bars show 2B and darker bars show 9B, with identical score axes across panels. The 2B bars use unrounded source values, with labels rounded to one decimal. Headings report the seven-task mean at each scale.

Relation to training efficiency. These results complement the timing comparison for the third research question. In Table 1, the recorded mean step time grows from 31.79 to 182.18 seconds for T^5, compared with 54.01 to 497.83 seconds for RLP. Thus, the relative step-time reduction increases from 41.1% at 2B to 63.4% at 9B while T^5 retains gains over Base and NTP at the larger scale. Together, the measurements describe performance and per-step cost as model size grows, rather than time to a common target score.

A.7 HISTORICAL NAIVE-PPO DIAGNOSTIC

Figure 1 combines thought and mixed gains from an earlier 60-step naive PPO run, without T^5 or the new R^2 gate. It evaluates 256 examples each from CommonsenseQA (Talmor et al., 2019) and GSM8K (Cobbe et al., 2021) every ten steps. The actor is frozen through step 20. The retained record does not specify the allocation of these warmup steps between value heads and backbones. The evaluation seed is fixed within the run. Thought gain is $\ell_0 - \ell_{\text{thought}}$, while mixed gain substitutes the learned prediction mixture's loss. The no-thought prediction is therefore the zero-gain reference by construction. In this sense, the figure is an ablation-style diagnostic of prediction mode at a fixed checkpoint, not a comparison between separately trained Base and RL models.

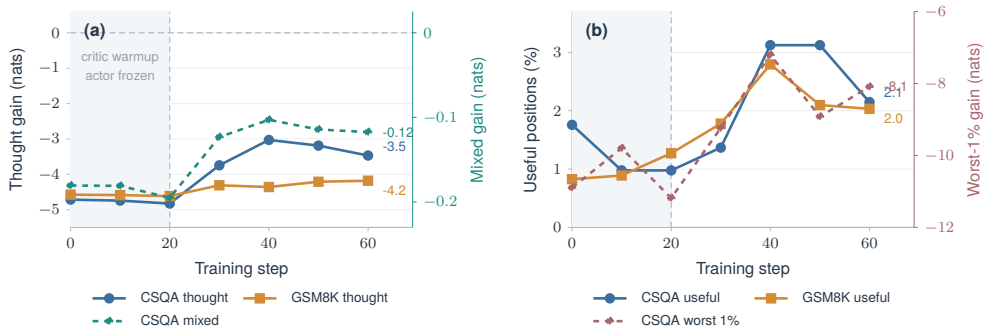


Figure 9: Historical naive-PPO ablation-style diagnostic at seven checkpoints. (a) Thought gain (left axis) and CSQA mixed gain (right axis). (b) Useful-position fraction (left axis) and CSQA worst-1% thought gain (right axis). The no-thought prediction has zero gain by construction. Shading marks the 20-step critic warmup with the actor frozen; endpoint labels are rounded values reconstructed from the original vector figure.

The redraw reconstructs the seven coordinates of each curve from the original vector PDF and its axis ticks. These are figure-derived values, not recovered raw run logs. Raw thought insertion remains below the zero reference on average, whereas prediction mixing substantially narrows the deficit and becomes slightly positive at the final checkpoint. This historical contrast motivates qualification and mixing but does not evaluate T^5. In the current experiments, T^5 exceeds Base on all seven reported tasks at both 2B and 9B. Extra stored decimal places preserve geometry and do not imply measurement precision. No multi-seed statistics or causal attribution to critic staleness are inferred from this diagnostic. Source assets, extraction metadata, and the plotting scripts are retained with the manuscript materials.

B WHY PREDICTIVE ACCURACY DOES NOT REMOVE DRIFT

Predictive qualification and conditional calibration address different parts of the update. We first separate policy reuse from value error, relate critic accuracy to advantage perturbations, and then show why accurate global prediction can coexist with harmful local drift. This motivates the conditional-moment objective proved in Appendix C.

B.1 WHAT CHANGES WHEN OLD ROLLOUTS TRAIN A NEW POLICY?

This subsection derives the distinction used in Section 3.2: policy lag changes the trajectories relevant to the objective, probability mismatch changes the correctness of recorded ratios, and value error changes the advantage estimate. The derivations concern a finite undiscounted horizon, a fixed distribution of text items, fixed transitions, and a fixed scorer and reward normalizer. Variable-length thoughts may be padded after termination with zero rewards and a deterministic dummy action. All policy gradients below assume differentiable parameter-independent support and sufficient integrability to exchange derivatives and expectations. The time index is included in the state.

For completeness, the return and ideal policy-gradient definitions used in the main text are

$$G_t = \sum_{\ell=0}^{T-t} \gamma^\ell r_{t+\ell}. \tag{16}$$

For the undiscounted case $\gamma = 1$, let τ denote a complete thought trajectory, including its text item, and write $u_t = \nabla_\theta \log \pi_\theta(a_t \mid s_t)$. With scoring and transitions fixed, the expected-return objective and its policy gradient are

$$J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}[G_1], \qquad \nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_t G_t u_t\right]. \tag{17}$$

B.1.1 A FIXED REWARD, BUT A CHANGING DISTRIBUTION OF THOUGHTS

Let τ contain the text item and the complete sampled thought. Its probability factors into the text distribution, transitions, and action probabilities. Only the last depends on θ. Write $G_1 = \sum_{k=1}^{T} r_k$, $J(\theta) = \mathbb{E}_{\pi_\theta} G_1$, and $u_t = \nabla_\theta \log \pi_\theta(a_t \mid s_t)$.

Lemma 1 (Policy gradient from trajectory likelihood)**.** *Under the fixed-reward conditions above,*

$$\nabla_\theta J = \mathbb{E}_{\pi_\theta}\left[G_1 \sum_t u_t\right] = \mathbb{E}_{\pi_\theta}\left[\sum_t G_t u_t\right]. \tag{18}$$

Proof. Write $p_\theta(\tau)$ for the trajectory probability and differentiate the expectation:

$$\nabla_\theta \sum_\tau p_\theta(\tau) G_1(\tau) = \sum_\tau p_\theta(\tau) G_1(\tau) \nabla_\theta \log p_\theta(\tau), \tag{19}$$

$$\nabla_\theta \log p_\theta(\tau) = \sum_t \nabla_\theta \log \pi_\theta(a_t \mid s_t). \tag{20}$$

This proves the first expression. For the second, separate rewards before action t from rewards at and after it. Conditional on the history before a_t, earlier rewards are fixed, while

$$\mathbb{E}_{a_t\sim\pi_\theta}[u_t \mid s_t] = \sum_a \nabla_\theta \pi_\theta(a \mid s_t) = \nabla_\theta 1 = 0. \tag{21}$$

Earlier rewards therefore contribute zero. The term for action t retains only $G_t = \sum_{k=t}^{T} r_k$. □

This is why it is more precise to say that a return sample was generated under a policy than to label the score of a fixed trajectory as a different reward for each policy. The same thought receives the same score within the window. What changes is its likelihood, and the expected quality of continuations from each partial thought.

Now let q be the actual behavior law, fixed while differentiating the actor. Naively replaying old trajectories gives

$$g_{\text{naive}}(\theta) = \mathbb{E}_{\tau\sim q}\left[\sum_t G_t u_t\right], \tag{22}$$

which differs from Eq. (18) through the trajectory distribution. If q covers the actor's trajectory support, the full likelihood ratio is

$$W_\theta(\tau) = \prod_{t=1}^{T} \frac{\pi_\theta(a_t \mid s_t)}{q(a_t \mid s_t)}. \tag{23}$$

The initial text distribution and transitions cancel. Provided the weighted expectation is integrable, multiplying each trajectory by this ratio restores the exact gradient:

$$\mathbb{E}_q\left[W_\theta(\tau)\sum_t G_t u_t\right] = \sum_\tau p_q(\tau)\frac{p_\theta(\tau)}{p_q(\tau)}\sum_t G_t u_t = \nabla_\theta J(\theta). \tag{24}$$

Long products can concentrate weight on a few trajectories even when individual token ratios are moderate. This variance cost motivates local surrogates and bounded reuse, rather than assuming that token-level reweighting fully corrects a sequence distribution.

B.1.2 Why an old-policy advantage is intentional in PPO

For this comparison assume $q = \pi_{\text{old}}$ exactly, and let d_t^π be the distribution of prefixes before token t under policy π. Define $Q_t^\pi(s,a) = \mathbb{E}_\pi[G_t \mid s_t = s, a_t = a]$, $V_t^\pi(s) = \mathbb{E}_\pi[G_t \mid s_t = s]$, and $A_t^\pi = Q_t^\pi - V_t^\pi$, retaining the fixed-text conditioning convention. The ideal L_{old} in Eq. (25) uses exact, unnormalized A_t^{old} and every action in each old trajectory. PPO deliberately makes a local approximation: before clipping and normalization, its ideal old-policy surrogate is

$$L_{\text{old}}(\theta) = \mathbb{E}_{\tau\sim\pi_{\text{old}}}\left[\sum_t \rho_t A_t^{\text{old}}\right], \qquad \nabla_\theta L_{\text{old}}(\theta)|_{\theta=\theta_{\text{old}}} = \nabla_\theta J(\theta_{\text{old}}). \tag{25}$$

Proposition 2 (The old-policy surrogate matches the gradient at its origin)**.** *For the fixed-reward, full-trajectory setting above,*

$$\nabla_\theta L_{\text{old}}(\theta)|_{\theta=\theta_{\text{old}}} = \nabla_\theta J(\theta_{\text{old}}). \tag{26}$$

Proof. At the old parameter, $\rho_t = 1$ and $\nabla_\theta \rho_t = u_t$. The state value contributes zero by the conditional score identity, hence

$$\nabla_\theta L_{\text{old}}|_{\theta_{\text{old}}} = \mathbb{E}_{\text{old}}\sum_t A_t^{\text{old}} u_t = \mathbb{E}_{\text{old}}\sum_t Q_t^{\text{old}} u_t. \tag{27}$$

The factor u_t is fixed after conditioning on (s_t, a_t), so averaging $G_t u_t$ over the remaining continuation gives $Q_t^{\text{old}} u_t$. Reversing this conditional expectation and applying Lemma 1 proves the result. □

Away from this starting point, exact current-action reweighting leaves the following two expressions:

$$\nabla_\theta L_{\text{old}}(\theta) = \sum_t \mathbb{E}_{s\sim d_t^{\text{old}},\, a\sim\pi_\theta}[A_t^{\text{old}}(s,a) u_t], \tag{28}$$

$$\nabla_\theta J(\theta) = \sum_t \mathbb{E}_{s\sim d_t^{\pi_\theta},\, a\sim\pi_\theta}[A_t^{\pi_\theta}(s,a) u_t]. \tag{29}$$

The first expression keeps old prefixes and old continuation values. The second uses the prefixes and continuations induced by the current actor. A token ratio changes neither of those other factors. Thus old-policy advantages are the reference for a local policy improvement, not evidence that the critic must already predict the updated policy perfectly. This is the standard surrogate distinction underlying TRPO and PPO (Schulman et al., 2015; 2017).

The identity at the origin concerns exact old-policy advantages (or unbiased Monte Carlo estimates of them) and full-trajectory summation. Normalized GAE, action-dependent mixture weights, and clipping away from the origin change that estimator. Section 5.2 controls a component of the actual mixed PPO gradient, not an exact ∇J. Relating them requires matching the occupation measure and time weights. Appendix E.4 fixes the aggregation for the auxiliary objective.

B.1.3 VALUE-LEARNING ERROR AND A MOVING VALUE TARGET

Fix a prefix and scoring item. Adding and subtracting V^q gives the decomposition in Section 4.1. The first term compares the fitted critic with the current rollout return target. Its fitting data may come from an earlier distribution. The second compares two different continuation policies, even if the first prediction were exact. For example, suppose a remaining binary decision returns one on success and zero otherwise. If the old policy succeeds with probability 0.2 and the new policy with probability 0.8, a perfect old critic predicts 0.2. Its difference from the new value is -0.6, entirely due to policy change, not failed old-policy fitting. More regression on old returns keeps the optimum at 0.2.

To expose historical target lag, let q_β denote one earlier continuation law under the same fixed reward. Then

$$V_\psi - V^q = (V_\psi - V^{q_\beta}) + (V^{q_\beta} - V^q). \tag{30}$$

The first term is fitting error relative to that earlier target, and the second is target lag. A critic trained on mixed history need not equal the value of any single policy. q_β illustrates one source rather than identifying an exact critic age. Scorer or reward-scale changes introduce further target changes and are excluded from this fixed-reward comparison.

There is an additional information distinction in this paper. Let W be the observed future text used for scoring, and let $V^q(s, W) = \mathbb{E}_q[G_t \mid s, W]$. Under ordinary squared regression, an unrestricted predictor that sees only s is minimized at

$$\bar{V}^q(s) = \mathbb{E}_{W|s}[V^q(s, W)]. \tag{31}$$

Indeed, expanding around $\bar{V}^q(s)$ makes the conditional regression risk a variance term plus $(V_i(s) - \bar{V}^q(s))^2$. The item-level error further separates as

$$V_i(s) - V^q(s, W) = [V_i(s) - \bar{V}^q(s)] + [\bar{V}^q(s) - V^q(s, W)]. \tag{32}$$

Only the first part is visible-state fitting error. The second reflects information withheld from the critic. Appendix D.2 quantifies its variance floor. Predictive qualification consequently checks useful out-of-sample return structure, not exact reconstruction of every item's oracle value.

B.1.4 BASELINE CANCELLATION UNDER EXACT ACTION RATIOS

Freeze a state-only baseline $b(s)$ before drawing the evaluated action and continuation. At any fixed prefix, assume exact action ratios $\rho = \pi_\theta / q$ and full actor-support coverage. Then, regardless of how different the actor is from q,

$$\mathbb{E}_q[\rho b(s) u_t \mid s] = b(s) \sum_a \pi_\theta(a \mid s) \nabla_\theta \log \pi_\theta(a \mid s) = 0. \tag{33}$$

The baseline need not equal either V^q or V^{π_θ}. This cancellation concerns the baseline contribution only, not the old-state and old-continuation errors in Eq. (29). Fitting a baseline on the evaluated actions and then detaching it is not equivalent to fixing it independently beforehand.

Two operations can break the cancellation. With an inexact recorded denominator, the remaining action weight is q/π_{old}. With clipping, a sample-dependent coefficient I selects which directions remain active. Even without clipping the first effect gives

$$\mathbb{E}_q[\rho u_t \mid s] = \sum_a \pi_\theta(a \mid s) \frac{q(a \mid s)}{\pi_{\text{old}}(a \mid s)} u_t, \tag{34}$$

which need not vanish. GAE also introduces future value errors rather than merely subtracting the current state baseline, as derived in Proposition 4. Section 5 addresses the resulting conditional-mean component with the actual mixed PPO mask, not every source of surrogate error.

These distinctions apply to both small and large RL problems. Long language-model continuations, sparse terminal signals, expensive rollout refresh, and seldom-repeated exact prefixes make policy lag and value generalization practically relevant. They do not imply that PPO is necessarily less reliable than GRPO. GRPO removes learned-value regression but still uses sampled advantages and must handle rollout reuse and recorded probabilities (Shao et al., 2024). Finally, our scorer is frozen only within a window. If one instead differentiated a changing reward $G_1(\theta)$ as part of the objective, the exact derivative would include $\mathbb{E}_{\pi_\theta}[\nabla_\theta G_1(\theta)]$ in addition to the score-function term. Our actor update treats scores as fixed. Updating the scorer starts a new objective window rather than differentiating through that reward term.

B.2 VALUE ERROR AND PREDICTIVE QUALIFICATION

This subsection explains why both critics are qualified separately. All expectations here use a fixed probability distribution of text items and valid-action prefixes, followed by continuations from q. Conditioning on s also fixes the scoring text, as in Section 3. The oracle V^q is the corresponding conditional return, whether or not a visible-prefix critic can represent it. Fix θ, the recorded denominator, scorer, critics, and normalizers. Assume differentiability, integrability of the gradients below, and finite $C_\rho^2 = \mathbb{E}_q[\rho^2 \|\nabla_\theta \log \pi_\theta(a \mid s)\|_2^2]$. Write $e_i = V_i - V^q$ and $u = \nabla_\theta \log \pi_\theta(a \mid s)$. Population predictive scores use this probability measure. Moving to an unnormalized occupation measure requires rescaling the gradients, risks, and C_ρ together.

B.2.1 CLIPPING DOES NOT AMPLIFY ADVANTAGE PERTURBATIONS

First consider shared-return Monte Carlo advantages $A_i = G - V_i(s)$. Compare a detached convex mixture, allowing $w = w(s,a) \in [0,1]$, with the oracle advantage $G - V^q(s)$ under the same normalizer.

Proposition 3 (Value error controls update error)**.**

$$\|g_{\text{twin}} - g_{\text{oracle}}\|_2 \le \frac{C_\rho}{\sigma_{\text{pilot}}} \sqrt{\mathbb{E}_s[e_1(s)^2 + e_2(s)^2]}. \tag{35}$$

For a scalar normalized advantage x, the per-sample clipped-PPO gradient is $\rho u H_\rho(x)$, where, away from clipping boundaries,

$$H_\rho(x) = \begin{cases} \max(x,0), & \rho < \rho_-, \\ x, & \rho_- < \rho < \rho_+, \\ \min(x,0), & \rho > \rho_+. \end{cases} \qquad |H_\rho(x) - H_\rho(y)| \le |x - y|. \tag{36}$$

At a clipping boundary use one fixed convex combination of the two adjacent branch maps. The same convention is used for the two advantages being compared.

Lemma 2 (The active advantage is nonexpansive)**.** *For any fixed ratio ρ and real numbers x, y,*

$$|H_\rho(x) - H_\rho(y)| \le |x - y|. \tag{37}$$

Proof. For $H(x) = \max(x, 0)$, two positive inputs retain their original distance, two negative inputs both become zero, and inputs on opposite sides of zero become closer because the negative part is removed. Thus their distance cannot increase. The proof for $\min(x, 0)$ is identical with signs reversed. The identity map preserves distance. A convex combination of such maps also obeys the bound by the triangle inequality. This covers the boundary convention. □

The lemma controls a switch of clipping branch without pretending that the two masks stay equal.

Proof of Proposition 3. For raw advantages A, A_o with the same normalizer, subtract their gradient contributions and move the norm inside the expectation:

$$\|g(A) - g(A_o)\|_2 \le \mathbb{E}_q[\rho \|u\|_2 |H_\rho(\widetilde{A}) - H_\rho(\widetilde{A}_o)|] \tag{38}$$

$$\le \frac{1}{\sigma_{\text{pilot}}} \mathbb{E}_q[\rho \|u\|_2 |A - A_o|] \tag{39}$$

$$\le \frac{C_\rho}{\sigma_{\text{pilot}}} \sqrt{\mathbb{E}_q[(A - A_o)^2]}. \tag{40}$$

The second line uses Lemma 2. The last uses Cauchy–Schwarz on the two nonnegative factors $\rho\|u\|_2$ and $|A - A_o|$.

For Monte Carlo advantages, subtracting the two baselines cancels the shared return: $A - A_o = -we_1 - (1-w)e_2$. The square of this mixture satisfies

$$we_1^2 + (1-w)e_2^2 - [we_1 + (1-w)e_2]^2 = w(1-w)(e_1 - e_2)^2 \ge 0. \tag{41}$$

Therefore $(A - A_o)^2 \le we_1^2 + (1-w)e_2^2 \le e_1^2 + e_2^2$. Insert this into Eq. (40) to obtain the proposition. □

The rollout distribution, clipping rule, and normalizer stay the same in this comparison. The pointwise convex-error inequality remains valid for a detached action-dependent weight. It does not invoke state-baseline cancellation. Expectations use the declared rollout occupation measure. Prediction risk under another distribution bounds actor-distribution error only with coverage control.

B.2.2 Predictive scores and independent validation

A collapsed predictor illustrates why the admission rule in Section 4.2 tests both critics separately. Consider equally represented held-out states $s_\pm$ with $G(s_\pm) = \pm r_\star$, $r_\star > 0$. Let $V_1(s_\pm) = \pm r_\star$ and $V_2(s_\pm) = 0$. Then

$$\mathrm{MSE}(V_1) = 0, \quad \mathrm{MSE}(V_2) = r_\star^2, \quad (R_1^2, R_2^2) = (1, 0). \tag{42}$$

As $r_\star \to 0$, the second critic's loss and twin disagreement vanish, although it has learned no return variation. Even the average predictor attains $R^2 = 3/4$. Testing the critics separately distinguishes this case from two qualified predictors, whereas a fixed warmup duration or small fitting loss does not. Let $\mathcal{E}_i = \mathbb{E}_q[(G - V_i(s))^2]$, $S_G^2 = \mathrm{Var}_q(G) > 0$, and $R_{i,\mathrm{pop}}^2 = 1 - \mathcal{E}_i / S_G^2$. All three quantities use the same distribution, with fixed critics and finite return moments.

Lemma 3 (Return error includes value error)**.** *For each critic,*

$$\mathcal{E}_i = \mathbb{E}_s[e_i(s)^2] + \mathbb{E}_s \mathrm{Var}_q(G \mid s). \tag{43}$$

Proof. Write $G - V_i = (G - V^q) - e_i$ and expand the square. Conditional on s, the cross term is $-2e_i \mathbb{E}_q[G - V^q \mid s] = 0$, because $V^q = \mathbb{E}_q[G \mid s]$. The remaining two terms are the conditional variance of G and e_i^2. Average them over states. □

Summing the identity for the two critics and substituting the definition of their population scores gives

$$\mathbb{E}_s[e_1^2 + e_2^2] = S_G^2(2 - R_{1,\mathrm{pop}}^2 - R_{2,\mathrm{pop}}^2) - 2\mathbb{E}_s \mathrm{Var}_q(G \mid s) \tag{44}$$

$$\leq S_G^2(2 - R_{1,\mathrm{pop}}^2 - R_{2,\mathrm{pop}}^2). \tag{45}$$

Return-prediction risk measures value error plus random variation among continuations. It is therefore conservative for Proposition 3. A global score controls an average, leaving room for the rare-state example in Appendix B.4.

The empirical R_i^2 in Eq. (5) is a diagnostic unless accompanied by generalization control. One sufficient construction works directly with squared residuals. For critics fixed before N_h independent holdout draws, suppose $|G - V_i(s)| \leq K_V$ on the evaluated distribution. With probability at least $1 - \delta$, simultaneously for both critics,

$$\mathcal{E}_i \leq \widehat{\mathcal{E}}_i + K_V^2 \sqrt{\frac{\log(2/\delta)}{2N_h}}, \qquad \widehat{\mathcal{E}}_i = \frac{1}{N_h} \sum_j (G_j - V_i(s_j))^2. \tag{46}$$

On this event, $\mathbb{E}_s(e_1^2 + e_2^2)$ is bounded by the sum of the two upper estimates. A gradient certificate additionally requires a bound on C_ρ. Correlated tokens from one trajectory are not independent holdout draws. Clustered validation and repeated qualification require appropriate concentration bounds and a declared total failure probability.

B.2.3 How value errors propagate through GAE

Compare learned and oracle GAEs on the same trajectory of realized length L, with the same $\lambda = \lambda(L)$, $0 \leq \gamma, \lambda \leq 1$, and zero terminal bootstrap. Let $\zeta_{i,t} = A_{i,t} - A_{o,t}$ and $w_k = \gamma(1-\lambda)(\gamma\lambda)^{k-1}$.

Proposition 4 (GAE propagates value error along the suffix)**.** *For each realized trajectory and pre-action position t,*

$$A_{i,t} - A_{o,t} = -e_i(s_t) + \gamma(1-\lambda) \sum_{k=1}^{L-t} (\gamma\lambda)^{k-1} e_i(s_{t+k}), \tag{47}$$

$$\zeta_{i,t} = -e_i(s_t) + \sum_{k=1}^{L-t} w_k e_i(s_{t+k}). \tag{48}$$

Proof. Subtract the oracle TD error from the learned TD error. The reward cancels, leaving $\gamma e_i(s_{t+1}) - e_i(s_t)$. Summing these differences with the same GAE weights gives

$$\zeta_{i,t} = \sum_{k=0}^{L-t} (\gamma\lambda)^k [\gamma e_i(s_{t+k+1}) - e_i(s_{t+k})]. \tag{49}$$

The coefficient of the current error is -1. For an intermediate error $e_i(s_{t+k})$, $k \geq 1$, its two occurrences have net coefficient $\gamma(\gamma\lambda)^{k-1} - (\gamma\lambda)^k = w_k$. The final error is zero by terminal bootstrapping. Collecting coefficients proves the identity. □

The current error enters directly. Later errors enter with nonnegative, decaying weights. Let $W_t = \sum_k w_k$. For $\gamma\lambda < 1$, its geometric sum is at most $\gamma(1-\lambda)/(1-\gamma\lambda) \leq 1$. If $\gamma\lambda = 1$, every w_k is zero. Cauchy–Schwarz applied to vectors $(1, \sqrt{w_1}, \ldots)$ and $(-e_i(s_t), \sqrt{w_1}e_i(s_{t+1}), \ldots)$ gives

$$\zeta_{i,t}^2 \leq (1 + W_t) \left[e_i(s_t)^2 + \sum_{k=1}^{L-t} w_k e_i(s_{t+k})^2 \right]. \tag{50}$$

For the convex mixture, the error is $w\zeta_{1,t} + (1-w)\zeta_{2,t}$, with squared magnitude at most $\zeta_{1,t}^2 + \zeta_{2,t}^2$. Applying Eq. (40) therefore yields

$$\|g_{\text{twin}}^{\text{GAE}} - g_{\text{oracle}}^{\text{GAE}}\|_2 \leq \frac{C_\rho}{\sigma_{\text{pilot}}} \sqrt{\mathbb{E}_q[\zeta_{1,t}^2 + \zeta_{2,t}^2]}. \tag{51}$$

Uniform value error $|e_i| \leq \epsilon_i$ on the evaluated prefixes gives $|\zeta_{i,t}| \leq 2\epsilon_i$. At $\lambda = 1$, only the current error remains. These calculations explain why general GAE also needs accurate later-prefix values. The oracle comparison uses matched GAE and does not assume that its conditional mean vanishes.

B.3 Two distinct consistency conditions for cancellation

The factors in $\Delta_\mu(s) = \mu_{\widetilde{A}}(s)\, \mu_v(s)$ concern different objects. *Sampling–update consistency* asks whether action reweighting recovers the actor's zero-mean score. *Advantage–sampling consistency* asks whether the advantage is centered under the law supplying its actions and returns. These are logically distinct conditions, not a claim of statistical independence between the factors.

Fix a prefix and its scoring item, and condition on all snapshots chosen independently of the evaluated continuation. Write $u = \nabla_\theta \log \pi_\theta(a \mid s)$, $\rho = \pi_\theta/\pi_{\text{old}}$, $\mu_{\widetilde{A}} = \mathbb{E}_q[\widetilde{A} \mid s]$, and $\mu_v = \mathbb{E}_q[\rho I u \mid s]$. Assume differentiable parameter-independent actor support covered by q, positive recorded probabilities, exchangeable differentiation and summation, and finite required moments. Normalization statistics μ_{pilot} and $\sigma_{\text{pilot}} > 0$ are fixed.

Proposition 5 (Two routes to a zero mean-induced term)**.** *Under these fixed-prefix conditions:*

1. *If $q = \pi_{\text{old}}$ and $I = 1$, then $\mu_v(s) = 0$ for any actor parameter satisfying the assumptions, without requiring $\pi_\theta = \pi_{\text{old}}$. The same result holds for any covering sampler q if the ratio is instead exactly π_θ/q.*
2. *The exact advantage $A^q(s,a) = Q^q(s,a) - V^q(s)$ has zero conditional mean under q. For a Monte Carlo estimate $A_i = G - V_i(s)$ with $V^q(s) = \mathbb{E}_q[G \mid s]$, its normalized conditional mean is*

$$\mu_{\widetilde{A}_i}(s) = \frac{V^q(s) - V_i(s) - \mu_{\text{pilot}}}{\sigma_{\text{pilot}}}. \tag{52}$$

Thus $V_i = V^q$ and $\mu_{\text{pilot}} = 0$ give $\mu_{\widetilde{A}_i}(s) = 0$, regardless of the actor ratio or clipping mask.

Consequently, either $\mu_v(s) = 0$ or $\mu_{\widetilde{A}}(s) = 0$ suffices for $\Delta_\mu(s) = 0$. Both need not hold.

Proof. For the first part, the exact ratio changes the action measure, leaving

$$\mu_v(s) = \sum_a q(a \mid s) \frac{\pi_\theta(a \mid s)}{q(a \mid s)} \nabla_\theta \log \pi_\theta(a \mid s) \tag{53}$$

$$= \sum_a \pi_\theta(a \mid s) \nabla_\theta \log \pi_\theta(a \mid s) = \nabla_\theta \sum_a \pi_\theta(a \mid s) = 0. \tag{54}$$

This is the score-function identity, not an equality-of-policies assumption. For the second part, define $Q^q(s,a) = \mathbb{E}_q[G \mid s,a]$. Iterated expectation gives $V^q(s) = \sum_a q(a \mid s) Q^q(s,a)$, so

$$\mathbb{E}_{a\sim q}[A^q(s,a) \mid s] = \sum_a q(a \mid s) Q^q(s,a) - V^q(s) = 0. \tag{55}$$

The sampled return residual $G - V^q(s)$ also has zero conditional mean, although it need not equal $A^q(s,a)$ on each continuation. Replacing V^q by the frozen V_i and applying the fixed affine normalizer yields Eq. (52). Finally, multiply the two factors in Eq. (8). □

Why practical PPO can have $\mu_v(s) \neq 0$. The active mask depends on the sampled advantage and hence may depend on the continuation, not just the action. Let $\bar{I}(s,a) = \mathbb{E}_q[I \mid s,a]$. Using the recorded ratio rather than assuming it is exact gives

$$\mu_v(s) = \sum_a \pi_\theta(a \mid s) \frac{q(a \mid s)}{\pi_{\text{old}}(a \mid s)} \bar{I}(s,a) u(s,a). \tag{56}$$

When $q = \pi_{\text{old}}$ and $I = 1$, all directions receive unit weight and cancel. An incorrect denominator or selective clipping can make the effective weights nonuniform and prevent cancellation. Neither necessarily makes μ_v nonzero: the weighted directions may still cancel. Thus $\mu_v = 0$ is not an if-and-only-if diagnostic of sampling–update consistency, and policy lag alone does not invalidate the first part of the proposition. Appendix C.8 bounds the effect of uneven weights. A zero population mean also does not make a finite replay's empirical score mean exactly zero.

What $\mu_{\widetilde{A}}(s)$ measures, and where the Monte Carlo interpretation ends. With $\mu_{\text{pilot}} = 0$, Eq. (52) identifies $\mu_{\widetilde{A}_i}$ as the scaled negative value error, $(V^q - V_i)/\sigma_{\text{pilot}}$. The value is conditional on the scoring item. A critic seeing only the visible prefix may not represent it exactly (Appendix B.1). For an exact advantage, iterated expectation also gives zero global population mean, so population normalization preserves conditional centering. An independently estimated and frozen pilot mean need not be zero: even an exact Monte Carlo critic then leaves $\mu_{\widetilde{A}_i} = -\mu_{\text{pilot}}/\sigma_{\text{pilot}}$. Recomputing statistics on the evaluated batch introduces dependence and does not justify treating them as fixed in this identity.

For GAE, value errors also enter through later prefixes, as Proposition 4 shows. Exact values with a fixed trace parameter give zero-mean oracle GAE by iterated expectation of the zero-mean TD residuals. A trace selected from the realized continuation length can correlate with those residuals, so exact values alone do not establish the same centering. Our length-adaptive implementation therefore estimates the conditional mean of the actual GAE signal rather than assuming it equals the current state's value error.

An inaccurate state baseline is harmless to the ideal, correctly reweighted and unclipped score gradient because $\mu_v = 0$, not because its advantage necessarily has $\mu_{\widetilde{A}} = 0$. Conversely, conditional centering eliminates the mean-induced term even if $\mu_v \neq 0$. T^5 targets the mixed GAE's conditional mean through the dual objective in Section 5. Action-dependent weights introduce the covariance in Eq. (61), so the mean is not a convex combination of two endpoint means. Its population risk is controlled only to the extent justified by the auxiliary approximation, optimization, and statistical errors in Theorem 1. Mixing changes its own PPO mask and can alter the covariance term. Eliminating Δ_μ neither recovers the complete current-policy gradient nor establishes baseline invariance under action-dependent weighting or clipping.

B.4 Why a fixed minimum does not calibrate advantages

We analyze minimum aggregation as a comparator to the calibrated convex weights in Section 5. This calculation does not attribute that implementation to the historical naive-PPO diagnostic. For two GAEs, let $\bar{A} = (A_1 + A_2)/2$ and $\Delta = |A_1 - A_2|/2$. Then

$$A_{\min} = \bar{A} - \Delta, \qquad \mu_{\widetilde{A}_{\min}}(s) = \frac{\mathbb{E}_q[\bar{A} \mid s] - \mathbb{E}_q[\Delta \mid s] - \mu_{\text{pilot}}}{\sigma_{\text{pilot}}}. \tag{57}$$

The minimum moves the conditional mean downward rather than selecting its distance to zero. For Monte Carlo advantages with zero terminal value,

$$A_{\min} = G - \max_i V_i(s), \qquad \mu_{\widetilde{A}_{\min}}(s) = \frac{V^q(s) - \max_i V_i(s) - \mu_{\text{pilot}}}{\sigma_{\text{pilot}}}. \tag{58}$$

A constant offset disappears if normalization is recomputed from that shifted batch, but a state-dependent offset need not. With a frozen normalizer, neither disappears automatically. These identities explain the limitation of pessimistic aggregation. The proposed method learns a frozen action-dependent rule $w_\varphi(s, a)$ that combines the two GAEs from each shared trajectory.

A harmful update with exact rollout probabilities. Let two actions return 1 and 0, with $q = \pi_{\text{old}}$ assigning each probability $1/2$. Set $V_1 = 3/2$, $V_2 = -1/2$, and $(\mu_{\text{pilot}}, \sigma_{\text{pilot}}) = (0, 1)$. Minimum aggregation gives advantages $(-1/2, -3/2)$. Parameterize the Bernoulli actor by $\pi_\theta(a_+) = 1/(1 + e^{-\theta})$ and take $\theta = \log(7/3)$, so $\pi_\theta(a_+) = 0.7$. For clipping bounds $[0.8, 1.2]$, ratios are $(1.4, 0.6)$: the rewarding action remains active because its advantage is negative, while the other action is clipped. Hence

$$\partial_\theta \mathcal{J}_{\text{clip}}(\theta \mid s) = \tfrac{1}{2}(1.4)(-\tfrac{1}{2})(0.3) = -0.105, \qquad \partial_\theta \mathbb{E}_{\pi_\theta}[G \mid s] = 0.21. \tag{59}$$

The active factors are $v_+ = 0.42$ and $v_- = 0$, so $\mu_{\widetilde{A}_{\min}} = -1$, $\mu_{v_{\min}} = 0.21$, and $\Delta_{\mu,\min} = -0.21$. The covariance is 0.105. The offset overwhelms this positive signal. The individual conditional offsets are -1 and 1, so weight $1/2$ removes this mean-induced component, but does not undo clipping of the remaining gradient. This is a constructed example, not a measured result.

High global scores can conceal the local error. To see why a global predictive gate need not exclude this example, let a fraction ω of a held-out set come from this state, with its two actions equally represented. Split the remaining samples equally between two deterministic-return states with $G = 1$ and $G = -1$, where both critics are exact. Each critic then has mean-squared error $5\omega/4$, while the held-out return variance is $1 - \omega/2 - \omega^2/4$. Consequently,

$$R_1^2 = R_2^2 = 1 - \frac{5\omega/4}{1 - \omega/2 - \omega^2/4} \longrightarrow 1 \quad \text{as } \omega \to 0. \tag{60}$$

Both critics can therefore pass any fixed threshold below one while retaining the local harmful update. This is a gap between global prediction accuracy and state-conditional update quality, not evidence that either gate guarantees improvement on every state.

C Conditional-Moment Learning and the Drift Bound

The proof follows the construction in Section 5: the paired advantages define the correction family, the dual learns its conditional mean, and the retention constraint limits displacement from the reference. We then transfer the remaining mean risk to the actual PPO drift, quantify independent validation, and refine the sensitivity factor.

Fix the rollout law q, scoring rule, critic snapshots, common pilot normalizer, and the token occupation measure d. The complete context c precedes the current action and includes the scoring item and reward history. Expressions conditioned on s below abbreviate conditioning on this c. The weight $w_\varphi(s,a)$ uses the visible prefix and current action. The auxiliary $h(c)$ may additionally use the fixed scoring context. Both parameter sets are fixed before the trajectories on which population statements or validation bounds are evaluated. Assume the stated expectations exist.

C.1 Action dependence and the limits of twin disagreement

Let $\bar{w}_\varphi(s) = \mathbb{E}_q[w_\varphi \mid s]$ and $\Delta\widetilde{A} = \widetilde{A}_1 - \widetilde{A}_2$. Expanding $\widetilde{A}_\varphi = \widetilde{A}_2 + w_\varphi \Delta\widetilde{A}$ gives

$$\begin{aligned}\mathbb{E}_q[\widetilde{A}_\varphi \mid s] &= \mu_{\widetilde{A}_2} + \mathbb{E}_q[w_\varphi \Delta\widetilde{A} \mid s] \\ &= \bar{w}_\varphi \mu_{\widetilde{A}_1} + (1-\bar{w}_\varphi)\mu_{\widetilde{A}_2} + \mathrm{Cov}_q(w_\varphi, \Delta\widetilde{A} \mid s). \end{aligned} \tag{61}$$

The covariance is zero for a state-only weight, but need not vanish for a frozen action-dependent rule. Its magnitude is at most

$$|\mathrm{Cov}_q(w_\varphi, \Delta\widetilde{A} \mid s)| \le \sqrt{\mathrm{Var}_q(w_\varphi \mid s)\mathrm{Var}_q(\Delta\widetilde{A} \mid s)}. \tag{62}$$

Freezing parameters preserves this dependence. A rule that takes realized GAEs as input also depends on the sampled continuation, unlike the $w(s,a)$ architecture in Section 5.

Twin disagreement does not identify a common conditional offset. Adding the same state-dependent shift $b(s)$ to both normalized advantages leaves $\Delta\widetilde{A}$ unchanged while increasing both conditional means by $b(s)$. Relative disagreement alone cannot distinguish these cases. In the proposed objective, trajectory labels train the separate conditional-mean function. The pair determines the directions in which the gate can alter that signal. The dual lemma below holds for any square-integrable signal, so its validity is not evidence that two return critics estimate the mean more accurately than one. Such a comparison also depends on the correction family, data, model errors, and computation.

In the Monte Carlo case, $A_i = G - V_i(s)$, the critics share the same random return. Their difference $A_1 - A_2 = V_2(s) - V_1(s)$ reveals relative prediction differences, not the common value error. It is state-only, so the covariance in Eq. (61) vanishes in this special case even for an action-dependent weight. The mean uses $\bar{w}_\varphi(s)$. General GAE instead includes random later-prefix value differences. Random initialization alone does not make two advantage outputs independent conditional draws from the rollout law. Action-dependent mixing can also turn a state baseline into an action-dependent quantity. With $b_i = V_i + \mu_{\mathrm{pilot}}$,

$$\widetilde{A}_\varphi = \frac{G - b_\varphi(s,a)}{\sigma_{\mathrm{pilot}}}, \qquad b_\varphi(s,a) = w_\varphi(s,a)b_1(s) + (1 - w_\varphi(s,a))b_2(s). \tag{63}$$

The usual state-baseline cancellation does not apply to $b_\varphi(s,a)$. Under an exact, unclipped on-policy score, its contribution contains $-(b_1 - b_2)\mathbb{E}_q[w_\varphi \nabla_\theta \log \pi_\theta \mid s]/\sigma_{\mathrm{pilot}}$. The method therefore does not claim baseline invariance or oracle policy-gradient recovery.

A centered mixture with nonzero signal. The before–after illustration in Eq. (10) uses equally likely actions a_+, a_-. Let $\widetilde{A}_1 = (0.9, -0.1)$ and $\widetilde{A}_2 = (0.3, -0.7)$, giving the fixed average $\overline{A} = (0.6, -0.4)$ and conditional mean 0.1. Choosing

$$w(a_+) = \tfrac{1}{2}, \qquad w(a_-) = \tfrac{1}{6} \quad \Longrightarrow \quad \widetilde{A}_w = (0.6, -0.6), \qquad \mu_{\widetilde{A}_w} = 0 \tag{64}$$

keeps both signs unchanged, so the PPO mask and $\mu_v(s)$ are unchanged at fixed policy ratios. The mean-induced component falls from $0.1\mu_v(s)$ to zero, while the action contrast increases from 1 to 1.2. The signal-retention quantities are $C = \frac{1}{2}(0^2 + (-0.2)^2) = 0.02$ and $Q = \frac{1}{2}(0.6^2 + (-0.4)^2) = 0.26$. Thus $C/Q = 1/13$, making this constructed mixture feasible for $\kappa \ge 1/13$ within the required range $\kappa < 1$. These are illustrative values, not an experimental setting.

C.2 THE CONDITIONAL-MOMENT DUAL

Recall $R(\varphi) = \mathbb{E}_d[\mu_{\widetilde{A}_\varphi}(s)^2]$. For a mean function $h(c)$ with finite $\mathbb{E}_d[h^2]$, the conditional-moment objective is

$$\mathcal{L}(\varphi, h) = \mathbb{E}_{d,q}[2h(c)\widetilde{A}_\varphi - h(c)^2]. \tag{65}$$

Lemma 4 (A scalar dual represents the conditional mean square)**.** *For any square-integrable mixed advantage and $h(c) \in L^2(d)$,*

$$\mathcal{L}(\varphi, h) = R(\varphi) - \|h - \mu_{\widetilde{A}_\varphi}\|^2_{L^2(d)}. \tag{66}$$

Consequently $\sup_{h \in L^2(d)} \mathcal{L}(\varphi, h) = R(\varphi)$*, attained at the conditional mean.*

Proof. Since h is measurable before the action, iterated expectation gives $\mathbb{E}_{d,q}[h(c)\widetilde{A}_\varphi] = \mathbb{E}_d[h(c)\mu_{\widetilde{A}_\varphi}(s)]$. Complete the square in $2h\mu - h^2 = \mu^2 - (h-\mu)^2$ and average. Square integrability makes the conditional mean an admissible maximizer. □

Writing $\widetilde{A}_\varphi = \mu_{\widetilde{A}_\varphi} + (\widetilde{A}_\varphi - \mu_{\widetilde{A}_\varphi})$ and using the zero conditional mean of the second term gives

$$\mathbb{E}_{d,q}[\widetilde{A}^2_\varphi] = R(\varphi) + \underbrace{\mathbb{E}_d \mathrm{Var}_q(\widetilde{A}_\varphi \mid s)}_{\text{within-context variation}}. \tag{67}$$

The auxiliary objective isolates a population conditional mean without paired continuations at each state, not an exact mean from one observed advantage. Fitting one label per context can memorize noise, so population evaluation remains necessary.

For a fixed mixture and $h_\zeta \in \mathcal{H} \subseteq L^2(d)$, define the restricted optimum $R_{\mathcal{H}}(\varphi) = \sup_{h \in \mathcal{H}} \mathcal{L}(\varphi, h)$, approximation error $\epsilon_{\mathrm{app}}(\varphi) = \inf_{h \in \mathcal{H}} \mathbb{E}_d[(h - \mu_{\widetilde{A}_\varphi})^2]$, and inner optimization gap $\epsilon_{\mathrm{opt}} = R_{\mathcal{H}}(\varphi) - \mathcal{L}(\varphi, h_\zeta)$. The dual identity gives $R_{\mathcal{H}} = R - \epsilon_{\mathrm{app}}$ and the decomposition

$$\epsilon_h := \mathbb{E}_d[(h_\zeta(c) - \mu_{\widetilde{A}_\varphi}(s))^2] = \epsilon_{\mathrm{app}} + \epsilon_{\mathrm{opt}}. \tag{68}$$

This is error against the conditional mean. Error against a sampled advantage additionally contains $\mathbb{E}_d \mathrm{Var}_q(\widetilde{A}_\varphi \mid s)$, so an empirical sample-regression loss is not the same quantity. The optimization gap concerns the auxiliary head at fixed mixture, not optimization of the gate.

If the auxiliary input is only a compressed feature $z(c)$, the unrestricted optimum on that input is $\mathbb{E}[\widetilde{A}_\varphi \mid z]$. The unresolved risk is exactly

$$R(\varphi) - \mathbb{E}_d[(\mathbb{E}_{d,q}[\widetilde{A}_\varphi \mid z])^2] = \mathbb{E}_d \mathrm{Var}_d(\mu_{\widetilde{A}_\varphi}(s) \mid z). \tag{69}$$

Conditional means and variances here use the probability law obtained by normalizing d. The outer integrals retain its original mass. The identity therefore also holds for the finite occupation measure. This identity follows by conditioning μ on z and expanding its second moment. In particular, omitting the scoring item can hide item-specific offsets rather than eliminate them.

C.3 ONE-ROLLOUT STOCHASTIC GRADIENTS

Keep q, critic features, and GAE endpoints detached during head fitting. Substituting the mixture into the objective gives

$$\mathcal{L}(\varphi, h) = \mathbb{E}_{d,q}[2h(c)\widetilde{A}_2 - h(c)^2] + 2\mathbb{E}_{d,q}[h(c)w_\varphi(s,a)\Delta\widetilde{A}]. \tag{70}$$

For differentiable heads, let $\ell_{\mathrm{mom}} = 2h_\zeta(c)\widetilde{A}_\varphi - h_\zeta(c)^2$. Its per-sample gradients are

$$\begin{aligned} \nabla_\varphi \ell_{\mathrm{mom}} &= 2h_\zeta(c)\Delta\widetilde{A}\,\nabla_\varphi w_\varphi(s,a), \\ \nabla_\zeta \ell_{\mathrm{mom}} &= 2(\widetilde{A}_\varphi - h_\zeta(c))\nabla_\zeta h_\zeta(c). \end{aligned} \tag{71}$$

The predicted offset sets the correction direction, and the advantage difference determines the available adjustment. If the advantages coincide, the weight gradient vanishes even when the predicted

mean is nonzero. Giving h the sampled action would change its target to an action-conditional mean and could suppress useful action differences. Assume these gradients and the constraint gradient are integrable and differentiation can be interchanged with expectation. A sufficient local condition is a uniformly bounded gate Jacobian and a common L^2 envelope for the auxiliary output and its Jacobian. The moment gradients are then unbiased for $\mathcal{L}$ at fixed parameters, while their relation to R depends on mean-approximation quality. Under these conditions,

$$\nabla_\varphi \mathcal{L} = \mathbb{E}_{d,q}[\nabla_\varphi \ell_{\mathrm{mom}}], \tag{72}$$

$$\nabla_\zeta \mathcal{L} = \mathbb{E}_{d,q}[\nabla_\zeta \ell_{\mathrm{mom}}], \tag{73}$$

$$\nabla_\varphi C = \mathbb{E}_{d,q}[2(\widetilde{A}_\varphi - \overline{A})\Delta\widetilde{A}\,\nabla_\varphi w_\varphi(s,a)]. \tag{74}$$

One complete trajectory gives unbiased trajectory-summed estimates of these expressions with the fixed aggregation in Appendix E.4. Tokens on a trajectory may be correlated, since linearity of expectation is sufficient for this statement. Independence is needed across fresh trajectories for the validation concentration bound below. Conditional on training history, an online update uses parameters fixed before its fresh trajectory is drawn. Reusing a fixed trajectory set gives stochastic gradients of its empirical objective. Evaluating the selected candidate on that same set does not provide independent population validation. This result concerns a stochastic observation, not the total number of trajectories needed to fit h or reduce R to a specified tolerance. Appendix E.7 accounts for every sampling phase.

For a gate function w, write $R(w), C(w), \mathcal{L}(w,h)$ for the same objectives evaluated at that function. Introducing the nonnegative retention multiplier β gives

$$\mathcal{J}(w,h,\beta) = \mathcal{L}(w,h) + \beta\{C(w) - \kappa Q\}. \tag{75}$$

A practical update evaluates Eq. (75) at $w = w_\varphi$ and $h = h_\zeta$, descending in φ and ascending in ζ, β. Its per-token integrand is

$$j = 2h_\zeta(c)\widetilde{A}_\varphi - h_\zeta(c)^2 + \beta\{(w_\varphi - \tfrac{1}{2})^2(\Delta\widetilde{A})^2 - \kappa\overline{A}^2\}. \tag{76}$$

The weight gradient adds $\beta\nabla_\varphi C$ to the moment gradient, and the multiplier derivative is $(w_\varphi - \frac{1}{2})^2(\Delta\widetilde{A})^2 - \kappa\overline{A}^2$. All products use the two advantages from the same trajectory. No independent continuation is needed to form a product or a prefix-wise Monte Carlo mean. Projected ascent keeps β nonnegative. Proposition 1 characterizes the ideal function-space problem, not global convergence of neural alternating optimization. After fitting, freeze the candidate and use independent validation and actor trajectories. Detach the complete actor advantage so no actor gradient enters the heads, critics, or normalizer.

C.4 Population saddle points and optimal weights

Let $x = (s,a)$ denote only the visible gate input. Conditioning on x here does not invoke the shorthand that also fixes the hidden parts of c. Conditional expectations use the normalized joint law of (d,q), while outer integrals retain its finite nonzero mass M. Let $\mathcal{W}$ contain all measurable $[0,1]$-valued functions of x, and write $\widetilde{A}_w = \overline{A} + (w(x) - \frac{1}{2})\Delta\widetilde{A}$. The feasible set is $\mathcal{F} = \{w \in \mathcal{W} : C(w) \le \kappa Q\}$, and $R^* = \min_{w\in\mathcal{F}} R(w)$. The mixture is affine in w, so $\mathcal{L}$ is affine in w and concave in h. The quadratic C is convex, and the multiplier term is affine in β.

Under the assumptions of Proposition 1, $\mathcal{J}$ is convex in w and jointly concave in (h,β). We establish a saddle point with value R^* and derive the optimal gate in Eq. (14).

Proof of Proposition 1. Existence of a feasible optimum. Set $a(x) = \mathbb{E}_{d,q}[(\Delta\widetilde{A})^2 \mid x]$ and define $d\nu(x) = Ma(x)dP_x(x)$, where P_x is the marginal of the normalized joint law. Work in the Hilbert space $L^2(\nu)$. Gates that differ only where $a = 0$ have the same effect on the signal. The set $\mathcal{W}$ is nonempty, closed, bounded, and convex in this space, hence weakly compact. The linear operator $Tu = \mathbb{E}_{d,q}[u(x)\Delta\widetilde{A} \mid c]$ obeys, by conditional Jensen,

$$\|Tu\|^2_{L^2(d)} \le \mathbb{E}_{d,q}[u(x)^2(\Delta\widetilde{A})^2] = \|u\|^2_{L^2(\nu)}. \tag{77}$$

Thus $R(w) = \|\mu_{\overline{A}} + T(w - \frac{1}{2})\|^2_{L^2(d)}$ and $C(w) = \|w - \frac{1}{2}\|^2_{L^2(\nu)}$ are continuous convex functions and are weakly lower semicontinuous. The feasible set is a nonempty weakly compact subset of $\mathcal{W}$, so it contains a minimizer w^*.

Strong duality and the saddle. The reference satisfies $C(1/2) = 0 < \kappa Q$, giving Slater's condition relative to $\mathcal{W}$. Convex Lagrange duality therefore supplies a finite multiplier $\beta^* \geq 0$ such that w^* minimizes $R + \beta^* C$ over $\mathcal{W}$ and

$$C(w^*) \leq \kappa Q, \qquad \beta^*\{C(w^*) - \kappa Q\} = 0. \tag{78}$$

Take $h^* = \mathbb{E}_q[\widetilde{A}_{w^*} \mid c]$, the maximizing response from Lemma 4. For $u = w - w^*$, the directional derivative of $\mathcal{L}(\cdot, h^*)$ at w^* equals that of R. The first-order condition for $R + \beta^* C$, followed by expanding the quadratic constraint, gives

$$\begin{aligned}\mathcal{J}(w, h^*, \beta^*) - \mathcal{J}(w^*, h^*, \beta^*) = 2\mathbb{E}_{d,q}[h^* u \Delta\widetilde{A}] + 2\beta^* \mathbb{E}_{d,q}[(w^* - \tfrac{1}{2})u(\Delta\widetilde{A})^2] \\ + \beta^* \|u\|^2_{L^2(\nu)} \geq 0.\end{aligned} \tag{79}$$

The square-completion identity and complementary slackness give the other side. For all admissible w, h, β,

$$\mathcal{J}(w^*, h, \beta) \leq \mathcal{J}(w^*, h^*, \beta^*) = R^* \leq \mathcal{J}(w, h^*, \beta^*). \tag{80}$$

Consequently the unrestricted constrained problem has the strong-duality representation

$$\min_{w \in \mathcal{W}} \sup_{h \in L^2(d),\, \beta \geq 0} \mathcal{J}(w, h, \beta) = \max_{h \in L^2(d),\, \beta \geq 0} \min_{w \in \mathcal{W}} \mathcal{J}(w, h, \beta) = R^*. \tag{81}$$

For an infeasible gate the inner supremum is infinite through β, which is why the left side retains sup.

Optimal visible-input weights. Define $b_h(x) = \mathbb{E}_{d,q}[h(c)\Delta\widetilde{A} \mid x]$. Given h, β, the terms depending on $w(x)$ are

$$2b_h(x)(w(x) - \tfrac{1}{2}) + \beta a(x)(w(x) - \tfrac{1}{2})^2. \tag{82}$$

For $\beta > 0$ and $a(x) > 0$, its unconstrained derivative vanishes at $w = 1/2 - b_h/(\beta a)$. Projecting onto $[0, 1]$ and substituting h^*, β^* proves Eq. (14). Conditional Cauchy–Schwarz gives $|b_h|^2 \leq a\mathbb{E}[h^2 \mid x]$, so the ratio is well-defined on $a > 0$, and $b_h = 0$ on $a = 0$. The construction averages over scoring information and future randomness given visible x, so it respects the gate's input restriction. □

Boundary cases and interpretation. If $a(x) = 0$, both critics give identical advantages conditional on x almost surely, and any weight there is equivalent. If $\beta^* = 0$, minimize the linear term in Eq. (82): $b_{h^*} > 0$ requires $w^* = 0$, $b_{h^*} < 0$ requires $w^* = 1$, and $b_{h^*} = 0$ permits any weight consistent with the remaining optimality conditions. An active constraint need not have a positive multiplier. The relations are self-consistent because h^* depends on w^* and β^* satisfies Eq. (78). They do not prescribe substituting a realized GAE into the gate at actor-update time.

When $\kappa = 0$, feasibility fixes the mixed signal to $\overline{A}$, but strict feasibility fails and a finite saddle multiplier need not exist. The original signal-retention bound still applies. The case $Q = 0$ is treated in Appendix C.5. Neural parameterizations need not inherit convexity in φ, ζ, and empirical fitting remains distinct from the population problem. In particular, an unrestricted auxiliary function could memorize a single label at each training context, turning the empirical inner optimum into a sample-square loss. Shared representations and independent evaluation remain important. None of the optimality analysis adds a group-sampling step to the algorithm.

C.5 Signal retention and attainable risk

Proposition 6 (The displacement constraint excludes a zero signal)**.** *Let* $Q = \|\overline{A}\|^2_{L^2(d,q)} > 0$ *and* $C(\varphi) \leq \kappa Q$, $0 \leq \kappa < 1$. *Then*

$$\|\widetilde{A}_\varphi\|_{L^2} \geq (1 - \sqrt{\kappa})\sqrt{Q}, \qquad \mathbb{E}[\widetilde{A}_\varphi \overline{A}] \geq (1 - \sqrt{\kappa})Q. \tag{83}$$

Proof. Write $e = \widetilde{A}_\varphi - \overline{A}$. Using the common (d, q) weighting, the reverse triangle inequality gives the full chain

$$\|\widetilde{A}_\varphi\|_{L^2} \geq \|\overline{A}\|_{L^2} - \|\widetilde{A}_\varphi - \overline{A}\|_{L^2} = \sqrt{Q} - \sqrt{C(\varphi)} \geq (1 - \sqrt{\kappa})\sqrt{Q} > 0. \tag{84}$$

Cauchy–Schwarz gives $\mathbb{E}[\widetilde{A}_\varphi \overline{A}] = Q + \mathbb{E}[e\overline{A}] \geq Q - \sqrt{CQ}$. Insert $C \leq \kappa Q$ to obtain the alignment bound. □

These bounds preserve aggregate energy and alignment with the fixed twin mean, which need not be an oracle advantage. They do not preserve every state's action ordering, conditional covariance, or gradient direction. When $Q = 0$, the reference itself has no signal and feasibility forces a zero mixture almost surely. The window provides no policy-learning signal to retain. This case is diagnosed rather than interpreted as successful drift correction.

The constant gate $w = 1/2$ makes $C = 0$. Candidate weights are checked on independent complete trajectories. The base protocol falls back to this constant gate after a failed empirical check. That fallback guarantees feasibility, not improved drift. If population enforcement is claimed, a valid upper bound on $C - \kappa Q$ is required. Selecting or tuning κ on the same check invalidates a fixed-candidate confidence interpretation.

Proposition 7 (A population comparison with the fixed mean)**.** *Assume the gate family contains $w = 1/2$, denoted φ_0. A feasible global minimizer of the unrestricted dual satisfies $R(\varphi^*) \le R(\varphi_0)$. More generally, if $R_{\mathcal{H}}(\hat{\varphi}) \le R_{\mathcal{H}}(\varphi_0) + \epsilon_{\text{outer}}$, then*

$$R(\hat{\varphi}) \le R(\varphi_0) + \epsilon_{\text{app}}(\hat{\varphi}) - \epsilon_{\text{app}}(\varphi_0) + \epsilon_{\text{outer}}. \tag{85}$$

Proof. The reference is feasible. For the second statement substitute $R_{\mathcal{H}}(\varphi) = R(\varphi) - \epsilon_{\text{app}}(\varphi)$ on both sides of the assumed population objective comparison. □

This is a conditional comparison, not a claim that stochastic head training attains its premise. A nonzero feasible optimum is allowed. Shared critic bias, restricted inputs, and the retention constraint can all prevent exact centering.

C.6 Actual PPO gradient and error bounds

For the detached mixed advantage, use its own active mask

$$I_\varphi = \mathbf{1}\{\widetilde{A}_\varphi > 0, \rho < \rho_+\} + \mathbf{1}\{\widetilde{A}_\varphi < 0, \rho > \rho_-\} + \mathbf{1}\{\widetilde{A}_\varphi = 0\}, \qquad v_\varphi = \rho I_\varphi \nabla_\theta \log \pi_\theta. \tag{86}$$

At clipping boundaries choose a fixed subgradient coefficient in $[0, 1]$. The zero-advantage convention does not change its zero gradient contribution.

Clipping after mixing. For any scalar normalized advantage x, the clipped surrogate obeys the pointwise identity

$$\min\{\rho x, \text{clip}(\rho, \rho_-, \rho_+)x\} = \rho x - x_+(\rho - \rho_+)_+ - (-x)_+(\rho_- - \rho)_+. \tag{87}$$

The identity follows by checking the two signs of x and the clipping thresholds. It does not require a state-only weight. It also shows why clipping the two advantages separately and then mixing their gradients need not equal clipping their mixture. The conditional-moment objective trains the weight before actor sampling. It is not an additional differentiable penalty inside the actor loss.

For the actual mixed signal and its own mask,

$$g_\varphi(s) = \text{Cov}_q(\widetilde{A}_\varphi, v_\varphi \mid s) + \underbrace{\mu_{\widetilde{A}_\varphi}(s)\mu_{v_\varphi}(s)}_{\Delta_{\mu,\varphi}(s)}. \tag{88}$$

Expand the definition of conditional covariance. This step does not require independence between the weight and action or between the two factors. Detaching the advantage ensures that this is the score-form actor gradient, without derivatives through the heads.

Proof of Theorem 1. Using the actual mixed advantage and its own PPO mask, Cauchy–Schwarz gives

$$\mathbb{E}_d\|\Delta_{\mu,\varphi}\|_2 = \mathbb{E}_d[|\mu_{\widetilde{A}_\varphi}| \, \|\mu_{v_\varphi}\|_2] \tag{89}$$

$$\le \sqrt{R(\varphi)}\sqrt{\mathbb{E}_d\|\mu_{v_\varphi}\|_2^2} \le B\sqrt{R(\varphi)}. \tag{90}$$

Lemma 4 and Eq. (68) yield

$$R(\varphi) = \mathcal{L}(\varphi, h_\zeta) + \epsilon_h = \mathcal{L}(\varphi, h_\zeta) + \epsilon_{\text{app}} + \epsilon_{\text{opt}}. \tag{91}$$

Insert the independent validation upper bound on $\mathcal{L}$. This gives the theorem with ϵ_h, or equivalently with its approximation and inner-optimization components. On that event the sum under the square root is nonnegative, since it bounds $R \geq 0$. □

For all convex mixtures, a sufficient common factor is $B = C_\rho = (\mathbb{E}_{d,q}[\rho^2 \|\nabla_\theta \log \pi_\theta\|_2^2])^{1/2}$, by Jensen and $0 \leq I_\varphi \leq 1$. This factor must cover every actor parameter if the bound is claimed throughout PPO reuse. A reduced risk improves this common upper bound. It does not imply monotone decrease of the actual $\Delta_{\mu,\varphi}$ when its multiplier changes, or of full-gradient error. The clipping nonexpansiveness and GAE error bounds in Appendix B.2 remain separate comparisons with an oracle signal.

C.7 Independent trajectory validation and its limits

For a candidate (φ, ζ) fixed independently of n mutually independent validation trajectories drawn from the declared item and rollout law, use the trajectory statistic

$$Z_j = \frac{1}{T_{\max}} \sum_{t=1}^{T_{\max}} m_{jt}\{2h_\zeta(c_{jt})\widetilde{A}_{\varphi,jt} - h_\zeta(c_{jt})^2\}, \qquad \widehat{\mathcal{L}}_{\mathrm{val}} = \frac{1}{n}\sum_j Z_j, \tag{92}$$

where m_{jt} marks a valid pre-action token, as in Appendix E.4. If $|\widetilde{A}_i| \leq K$ and $|h_\zeta| \leq K_h$ are known population bounds, then $|Z_j| \leq C_L := 2K_hK + K_h^2$. Hoeffding's inequality gives, with probability at least $1-\delta$,

$$\mathcal{L}(\varphi, h_\zeta) \leq \widehat{\mathcal{L}}_{\mathrm{val}} + C_L\sqrt{\frac{2\log(1/\delta)}{n}}. \tag{93}$$

The independent units are trajectories, not their tokens. For the retention check, $0 \leq (\widetilde{A}_\varphi - \overline{A})^2 \leq K^2$ and $0 \leq \overline{A}^2 \leq K^2$. An upper confidence bound on $C - \kappa Q$ is its empirical trajectory average plus $(1+\kappa)K^2\sqrt{\log(1/\delta)/(2n)}$. Independent evaluation must follow candidate selection. Repeated checks require fresh trajectory sets and a declared error allocation.

Corollary 1 (What a drift certificate would require)**.** *If valid upper bounds a, o on $\epsilon_{\mathrm{app}}, \epsilon_{\mathrm{opt}}$ are also available, define*

$$U_R = \max\{0, \widehat{\mathcal{L}}_{\mathrm{val}} + \epsilon_{\mathrm{stat}} + a + o\}. \qquad U_R \leq \tau^2 \implies \mathbb{E}_d\|\Delta_{\mu,\varphi}\|_2 \leq B\tau. \tag{94}$$

The result is conditional on all these bounds. Concentration of Z_j alone does not upper-bound R, because a restricted or incompletely fitted auxiliary head provides a lower objective. A low empirical saddle value, a small empirical head gradient, or twin disagreement supplies neither a nor o. The base algorithm therefore uses empirical signal retention and reports the moment diagnostics without claiming this certificate. Bounded rewards do not automatically bound neural values or normalized GAE. Clipping only validation outputs would certify a different signal.

C.8 Policy sensitivity, clipping, and sampling mismatch

The coefficient B measures amplification by the actor and its active clipping branches. Fix one complete pre-action context and its frozen selected mixture, and write $\mu_{\widetilde{A}}, \mu_v, \Delta_\mu, I$ for the corresponding weighted quantities in this subsection. Let ξ denote randomness following the current action. Introduce

$$\nu_s(a, \xi) = \pi_\theta(a \mid s)q(\xi \mid s, a), \qquad \omega(a, s) = \frac{q(a \mid s)}{\pi_{\mathrm{old}}(a \mid s)}, \tag{95}$$

where ω is a sampling-mismatch ratio, not the mixture weight w. Only the current action is reweighted. The continuation law and both GAEs stay fixed. Assume defined continuations and positive recorded probabilities on actor support, differentiable parameter-independent actor support, and finite $F(s) = \mathbb{E}_{\pi_\theta}\|u\|_2^2$ and $\mathrm{Var}_{\nu_s}(\omega I)$, where $u = \nabla_\theta \log \pi_\theta(a \mid s)$.

Proposition 8 (Amplification by policy sensitivity and uneven weights)**.** *Under these fixed-context conditions,*

$$\|\Delta_\mu(s)\|_2 \leq |\mu_{\widetilde{A}}(s)|\sqrt{F(s)\mathrm{Var}_{\nu_s}(\omega I)}. \tag{96}$$

Proof. Expand the expectation by first sampling the action and then its continuation:

$$\mu_v(s) = \sum_a q(a \mid s) \frac{\pi_\theta(a \mid s)}{\pi_{\text{old}}(a \mid s)} \mathbb{E}_{\xi \sim q(\cdot|s,a)}[Iu] \tag{97}$$

$$= \sum_a \pi_\theta(a \mid s) \omega(a,s) \mathbb{E}_{\xi \sim q(\cdot|s,a)}[Iu] = \mathbb{E}_{\nu_s}[\omega I u]. \tag{98}$$

This changes only the action weighting. Since u depends on the current action, not on its future continuation, $\mathbb{E}_{\nu_s} u = \sum_a \nabla_\theta \pi_\theta(a \mid s) = 0$. Let $z = \omega I$ and $\bar{z} = \mathbb{E}_{\nu_s} z$. Subtracting this constant weight therefore leaves μ_v unchanged:

$$\mu_v = \mathbb{E}_{\nu_s}[(z - \bar{z})u], \qquad \|\mu_v\|_2 \le \mathbb{E}_{\nu_s}[|z - \bar{z}| \|u\|_2] \le \sqrt{\text{Var}_{\nu_s}(z) F(s)}. \tag{99}$$

The last step is Cauchy–Schwarz. Multiplying by $|\mu_{\widetilde{A}}|$ proves the proposition. The offset $\mu_{\widetilde{A}}$ itself is still evaluated under q. It was not redefined by this change of measure. □

For exact recorded probabilities on full actor support, $\omega = 1$. Away from clipping boundaries, I is binary. Write $p_c = \Pr_{\nu_s}(I = 0)$.

Corollary 2 (Selective clipping is the exact-ratio multiplier)**.** *With exact recorded probabilities and the binary mask above,*

$$\|\Delta_\mu(s)\|_2 \le |\mu_{\widetilde{A}}(s)| \sqrt{F(s) p_c(s)(1 - p_c(s))}. \tag{100}$$

Proof. A binary variable with mean $1 - p_c$ has second moment $1 - p_c$, so its variance is $(1 - p_c) - (1 - p_c)^2 = p_c(1 - p_c)$. Substitute $\omega = 1$ into Proposition 8. □

Applying Cauchy–Schwarz across the occupation measure also gives

$$\mathbb{E}_d \|\Delta_\mu\|_2 \le \sqrt{R(\varphi) \mathbb{E}_d[F p_c(1 - p_c)]}. \tag{101}$$

The three factors are offset size, policy sensitivity, and clipping imbalance. All branches active means score cancellation. All inactive means no update. Selectively removing directions can break the balance. At a boundary with $I \in [0, 1]$, use $p_c = 1 - \mathbb{E}_{\nu_s} I$ and $\text{Var}(I) \le \mathbb{E}I - (\mathbb{E}I)^2 = p_c(1 - p_c)$. Here p_c is a weight fraction rather than an event probability. In either case it uses current-action weighting, not the ordinary rollout clip fraction.

Tightness in the two-action example. At $\pi_\theta(a_+) = 0.7$, the Bernoulli scores are $u_+ = 0.3$ and $u_- = -0.7$. Only the positive-reward action remains active in Eq. (59), so

$$F = 0.7(0.3)^2 + 0.3(0.7)^2 = 0.21, \quad p_c = 0.3, \quad |\Delta_\mu| = 0.21 = |\mu_{\widetilde{A}}| \sqrt{F p_c(1 - p_c)}. \tag{102}$$

The example reaches equality because the surviving score direction aligns with the weight imbalance. Calibration changes the offset and can also change which branches survive. Both sides of the bound must use that selected mixture. Large negative-advantage ratios remain active, so clipping alone supplies no uniform bound on C_ρ or the mismatch factor. A guarantee over the entire update window requires a common bound on these factors. With no clipping and exact ratios, score cancellation already gives $\mu_v = 0$.

D SCOPE OF CONDITIONAL CALIBRATION

The preceding results concern the mean-induced term of the actual mixed GAE within a frozen window. We now examine what changes when the target is replaced by TD, scoring information is withheld, or the prefix distribution moves, before distinguishing neighboring actor–critic methods. These comparisons specify which assumptions must remain fixed for the guarantees to apply.

D.1 WHY ONE-STEP TD IS NOT SUBSTITUTED FOR GAE

Let $b_i(s) = \mathbb{E}_q[\delta_{i,t} \mid s_t = s]$ for a frozen critic. A single TD residual is an unbiased observation of

$$b_i(s) = (T^q V_i - V_i)(s), \tag{103}$$

where T^q includes reward and terminal masking. This is not generally the conditional mean of the GAE used by the actor. With a fixed trace parameter λ, a Markov context including time and reward history, and zero terminal bootstrap, define the raw mean $m_i(s) = \mathbb{E}_q[A_{i,t} \mid s_t = s]$. Iterated expectation gives

$$m_i(s) = \mathbb{E}_q[\delta_{i,t} + \gamma\lambda(1 - d_t)m_i(s_{t+1}) \mid s_t = s], \qquad \mu_{\widetilde{A}_i}(s) = \frac{m_i(s) - \mu_{\text{pilot}}}{\sigma_{\text{pilot}}}. \tag{104}$$

Replacing m_i by a learned bootstrap gives an unbiased sample of that model's Bellman target, not of the true m_i until its continuation mean is correct. For a fixed $w(s, a)$, the corresponding raw mixed mean is

$$\mathbb{E}_q[w\delta_1 + (1 - w)\delta_2 + \gamma\lambda(1 - d_t)\{wm_1(s_{t+1}) + (1 - w)m_2(s_{t+1})\} \mid s]\,. \tag{105}$$

The next-state terms retain the current weight. Substituting the next state's mixed mean is generally incorrect. The actual implementation uses $\lambda(L)$ selected by realized trajectory length. It can correlate with future residuals and cannot be pulled out of the conditional expectation to obtain Eq. (104) unchanged. The moment objective instead uses the actual sampled GAE and is valid with that length dependence. TD-based mean learning is a separate fixed-trace variant, not an unbiased shortcut asserted for the present algorithm.

D.2 TRAINING-TIME SCORING INFORMATION AND VISIBLE-STATE LIMITS

Let W denote the future-text scoring context and restore $V^q(s, W) = \mathbb{E}_q[G \mid s, W]$. Assume square-integrable conditional returns, a fixed $\sigma_{\text{pilot}} > 0$, and a square-integrable baseline $b(s)$ restricted to visible state.

Lemma 5 (Hidden scoring text leaves a variance floor)**.** *Under these information and moment conditions,*

$$\frac{\mathbb{E}_{s,W}[(V^q(s, W) - b(s))^2]}{\sigma^2_{\text{pilot}}} \geq \frac{\mathbb{E}_s \text{Var}_{W|s}(V^q(s, W))}{\sigma^2_{\text{pilot}}}. \tag{106}$$

Proof. Let $\bar{v}(s) = \mathbb{E}_{W|s} V^q(s, W)$ and write $V^q - b = (V^q - \bar{v}) + (\bar{v} - b)$. After squaring, the cross term has conditional mean zero since $\mathbb{E}_{W|s}(V^q - \bar{v}) = 0$. The remaining terms are $\text{Var}_{W|s}(V^q) + (\bar{v} - b)^2$. Drop the nonnegative square, average over s, and divide by σ^2_{pilot}. □

This floor applies directly to a state-only Monte Carlo baseline. An action-dependent $b_\varphi(s, a)$ is a different object and must be analyzed through the actual mixed conditional mean. The weight in Section 5 sees visible prefix/action features. The auxiliary mean function receives the complete conditioning information, including W, through a training-only input path. Neither the actor nor the return critics receive W. Restricting the auxiliary input to visible features instead targets an average over scoring items and leaves the information gap in Eq. (69). Even full context does not ensure representability by a finite auxiliary model, or existence of a feasible gate with zero offset.

D.3 WHY QUALIFICATION MUST FOLLOW THE PREFIX DISTRIBUTION

The learned gate defines a frozen rule on new prefixes, but its fitted occupation risk need not transfer to a different prefix distribution. The following supplementary result bounds this distribution effect while keeping the advantage-generating continuation law fixed. The occupancy argument follows trust-region analysis (Schulman et al., 2015). Consider fixed horizon T, identical initial text distributions and transitions, and policies q and π. Let d_t^q, d_t^π denote pre-action distributions, including the scoring item as an analytical context component. Keep both heads, critics, scorer, normalizer, and the continuation law defining $\mu_{\widetilde{A}_t}$ fixed. Assume this function is defined on both reachable supports and $|\mu_{\widetilde{A}_t}(s)| \le K$.

Define the time-averaged risk and the local action-distribution change by

$$\mathcal{L}_{d^\pi}(\mu_{\widetilde{A}}) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{d_t^\pi}[\mu_{\widetilde{A}_t}(s)^2], \qquad \delta_j = \mathbb{E}_{d_j^q} D_{\mathrm{TV}}(q(\cdot \mid s), \pi(\cdot \mid s)), \tag{107}$$

where $D_{\mathrm{TV}}(p,q) = \frac{1}{2}\|p-q\|_1$. It measures how much probability mass must be redistributed to turn one distribution into the other.

Proposition 9 (Early action changes affect more later prefixes)**.** *Under the fixed-horizon and frozen-score conditions above,*

$$\mathcal{L}_{d^\pi}(\mu_{\widetilde{A}}) \le \mathcal{L}_{d^q}(\mu_{\widetilde{A}}) + \frac{K^2}{T}\sum_{j=1}^{T-1}(T-j)\delta_j. \tag{108}$$

Proof. Step 1: separate old state differences from new action differences. Let P_q, P_π be the next-state transition kernels. Add and subtract the distribution obtained by applying π to the old states:

$$d_{t+1}^\pi - d_{t+1}^q = (d_t^\pi - d_t^q)P_\pi + d_t^q(P_\pi - P_q). \tag{109}$$

A common stochastic kernel cannot increase total variation: the triangle inequality and row sums of one give $\|(p-q)P\|_1 \le \sum_x |p(x)-q(x)| \sum_y P(y \mid x) = \|p-q\|_1$. Applying the same argument to the action-to-state transition bounds the second term by the average action difference δ_t. Therefore

$$D_{\mathrm{TV}}(d_{t+1}^\pi, d_{t+1}^q) \le D_{\mathrm{TV}}(d_t^\pi, d_t^q) + \delta_t \le \sum_{j=1}^{t}\delta_j, \tag{110}$$

since both policies start from the same initial distribution.

Step 2: turn state shift into score shift. For any function f in $[0, K^2]$, only the positive part of $p-q$ can increase its expectation, so $\mathbb{E}_p f - \mathbb{E}_q f \le K^2 \sum_{x:p(x)>q(x)}(p(x)-q(x)) = K^2 D_{\mathrm{TV}}(p,q)$. Apply this to $f = \mu_{\widetilde{A}_t}^2$.

Step 3: count how many times each action difference appears. Averaging over t gives a double sum. The change at position j appears for $t = j+1, \dots, T$, exactly $T-j$ times:

$$\frac{1}{T}\sum_{t=1}^{T} K^2 \sum_{j<t}\delta_j = \frac{K^2}{T}\sum_{j=1}^{T-1}(T-j)\delta_j. \tag{111}$$

This is the claimed bound. □

An early action can redirect more of the suffix, whereas the final action changes no later pre-action state. This counts possible propagation of state shift, not the reward importance of the last action. Pinsker's inequality also permits

$$\delta_j \le \sqrt{\tfrac{1}{2}\mathbb{E}_{d_j^q} D_{\mathrm{KL}}(q(\cdot \mid s) \,\|\, \pi(\cdot \mid s))}. \tag{112}$$

This relates bounded policy reuse to calibration coverage without changing the PPO objective. Both risks in Eq. (108) evaluate the same frozen $\mu_{\widetilde{A}_t}$, which still uses q continuations. Updating the continuation law, scorer, or normalizers changes that function and requires fresh validation. The displayed weights use fixed-length, uniform-time averaging. Variable-length valid-token normalization needs the corresponding occupancy weights rather than this formula unchanged.

D.4 Relation to Neighboring Actor–Critic Methods

SAO. Single-Rollout Asynchronous Optimization targets asynchronous agentic post-training, where each prompt produces one visible action trajectory and the learner may update while newer trajectories are still being generated (Hou et al., 2026). Its central problems are policy lag and the variance created by removing group-relative rollouts. SAO therefore forms token-level importance ratios against the rollout policy, masks ratios outside a double-sided interval, trains a single value model more frequently than the actor, freezes selected value-model parameters, and skips externally supplied observation tokens when propagating GAE.

T^5 operates on unlabeled text: hidden thought tokens are actions and teacher-forced continuation likelihood defines reward. There are no external observation steps to skip. Both methods use learned values and a single trajectory per selected item. Their learning targets and data schedules differ. T^5 separately qualifies two return critics, learns an action-dependent advantage mixture through a conditional-moment saddle objective, and freezes it for fresh actor trajectories. The scalar auxiliary mean function is distinct from a return critic. SAO and T^5 both use the length-adaptive GAE trace from VAPO (Yue et al., 2025). That component is not claimed as new.

TD3 and SAC. TD3 and SAC address continuous-control, off-policy actor–critic learning with replay buffers and action-value functions (Fujimoto et al., 2018; Haarnoja et al., 2018). TD3 reduces maximization bias with a minimum over bootstrapped Q targets and delays policy updates. SAC optimizes an entropy-regularized return through soft Bellman backups. T^5 instead uses locally on-policy collection with bounded within-iteration reuse, state-value critics, observed shaped returns, and a clipped likelihood-ratio objective over discrete latent tokens. Its action-dependent convex weights are learned through a conditional-moment objective, not a minimum over Bellman targets. The shared motivation is to manage function-approximation error. The mechanisms and data requirements differ. PPO, GAE, QUIET-STAR, and SAO are the more direct comparators.

E Training Protocol and Computational Cost

The analysis assumes a fixed sampling law, independent training and validation data, and detached advantages. The procedure below maintains these conditions within each training window and includes all sampling stages in the cost comparison.

E.1 Reward scale and checkpoint construction

The reward normalization uses an exponential second moment rather than subtracting a batch mean. Restoring the window index n, let

$$M_{r,n} = \beta M_{r,n-1} + (1-\beta)\frac{1}{|\mathcal{B}_n|}\sum_{\xi\in\mathcal{B}_n}\Psi_T(\xi)^2, \qquad \sigma_{r,n} = \sqrt{M_{r,n}+\varepsilon}, \tag{113}$$

where $0 \le \beta < 1$ and $\varepsilon > 0$. In the calibrated protocol, $\mathcal{B}_n$ contains historical or disjoint prior trajectories, not head-learning, validation, or actor trajectories from the current window. The resulting $\sigma_{r,n}$, abbreviated σ_r in the main text, remains fixed throughout that window.

Let $0 = \tau_0 < \tau_1 < \cdots < \tau_K = T$ be the scored checkpoints. Compute $\Psi_{\tau_k} = \ell_0 - \ell_{\tau_k}$ and $\Phi_{\tau_k} = \mathrm{clip}(\Psi_{\tau_k}/\sigma_r, -c_r, c_r)$, with $\Phi_0 = 0$. Set $\Phi_t = \Phi_{\tau_{k-1}}$ for $\tau_{k-1} < t < \tau_k$. Equation (2) is then equivalent to

$$r_t = \sum_{k=1}^{K} \mathbf{1}\{t = \tau_k\}(\Phi_{\tau_k} - \Phi_{\tau_{k-1}}). \tag{114}$$

The current length-12 configuration uses checkpoints 4, 8, and 12. In a variable-length rollout, the realized terminal prefix is included as the final checkpoint. One checkpoint gives a terminal-only shaped reward. It equals raw utility only without scaling and clipping.

Return conservation within a window. Fix one scorer and reward scale. Include the terminal prefix among checkpoints $0 = \tau_0 < \cdots < \tau_K = T$, set $\Phi_0 = 0$, and carry the last potential forward between checkpoints as in Eq. (2).

Proposition 10 (Checkpoint decomposition)**.** *Every thought trajectory satisfies*

$$\sum_{t=1}^{T} r_t = \Phi_T. \tag{115}$$

Proof. Unscored steps contribute zero. At scored steps, write the sum without cancelling terms first:

$$\sum_t r_t = (\Phi_{\tau_1} - \Phi_0) + (\Phi_{\tau_2} - \Phi_{\tau_1}) + \cdots + (\Phi_T - \Phi_{\tau_{K-1}}). \tag{116}$$

Every intermediate potential appears once with a plus sign and once with a minus sign. Only $\Phi_T - \Phi_0$ remains, and $\Phi_0 = 0$. □

With the raw potential, $\Phi_T = \ell_0 - \ell_T = \mathcal{U}$: intermediate scoring changes where credit is assigned, not its total. A common positive scale preserves ordering, while clipping merges utilities beyond its boundary. For example, any two thoughts with $\Psi_T > c_r\sigma_r$ both receive terminal potential c_r. Classical potential shaping has the form $r'(s,a,s') = r(s,a,s') + \gamma\Phi(s') - \Phi(s)$ and preserves optimal policies under fixed potentials and matching discount assumptions (Ng et al., 1999). Here the potential depends on the current language model and the scale in Eq. (113). We therefore use only the within-iteration telescoping identity. We do not claim stationary-MDP policy invariance across model updates.

E.2 Value regression and snapshot references

Here we restore version and parameter indices omitted in the main text. For critic $i \in \{1,2\}$, let V_{i,ψ_i} be the trainable model and $V_{i,k}$ its frozen pre-fit prediction in window k. With the return G_t in Eq. (16), the clipped regression objective is

$$\bar{V}_{i,\psi_i}(s_t) = V_{i,k}(s_t) + \mathrm{clip}(V_{i,\psi_i}(s_t) - V_{i,k}(s_t), -\epsilon_v, \epsilon_v)\,, \tag{117}$$

$$\mathcal{L}_{V_i}(\psi_i) = \frac{1}{2}\mathbb{E}_{s_t\sim\mathcal{D}_k^V}\left[\max\left\{(V_{i,\psi_i}(s_t) - G_t)^2, (\bar{V}_{i,\psi_i}(s_t) - G_t)^2\right\}\right]. \tag{118}$$

Critic-fitting outcomes are excluded from predictive holdout, pilot, head-learning, head-validation, and actor trajectory sets. Split complete text items or parent trajectories before extracting tokens, so correlated tokens are not treated as independent validation samples. Each selected text-position item contributes one ordinary trajectory in its assigned set, without requiring a bank of repeated-prefix continuations. Return targets retain their absolute shaped scale. The run configuration specifies whether full-critic warmup uses the clipped value loss above or ordinary half-squared error.

E.3 Advantage snapshots and masked GAE

Let $V_{i,k}^{\mathrm{adv}}$ denote the frozen snapshot used for actor advantages, distinct from the pre-fit value-clipping reference $V_{i,k}$. This is the post-regression model, frozen before pilot sampling and weight calibration and retained unchanged through head fitting and validation. The main text writes this snapshot as V_i when constructing actor advantages. With terminal indicator d_t and realized thought length L, the full recurrence is

$$\delta_{i,t} = r_t + \gamma(1 - d_t)V_{i,k}^{\mathrm{adv}}(s_{t+1}) - V_{i,k}^{\mathrm{adv}}(s_t), \tag{119}$$

$$A_{i,t} = \delta_{i,t} + \gamma\lambda(L)(1 - d_t)A_{i,t+1}, \qquad \lambda(L) = 1 - \frac{1}{\alpha L}. \tag{120}$$

The terminal bootstrap and subsequent advantage are zero, and α is chosen so $0 \le \lambda(L) \le 1$. We use $\gamma = 1$ for the actor and full-return targets for critic regression (equivalently $\lambda_V = 1$). The simplified Eq. (3) suppresses d_t under these terminal conventions.

Within window k, $\pi_{\mathrm{old}} = \pi_{\theta_k}$ is the recorded denominator, while q is the actual sampling law. On the valid thought tokens from an independent pilot trajectory set, pool the two raw GAE outputs with equal critic weight. Set μ_{pilot} to their pooled empirical mean and $\sigma_{\mathrm{pilot}} = \sqrt{\widehat{\sigma}^2_{\mathrm{pilot}} + \varepsilon_A}$ with $\varepsilon_A > 0$. Both channels use these same constants, which remain frozen during weight learning, validation, and PPO. They are not recomputed after selecting weights or within an actor batch. Other predetermined common pilot normalizers define a variant that must be reported separately.

E.4 Trajectory collection and actor replay

For each selected text-position item, sample one complete thought under the frozen rollout law q. Assign complete items or parent trajectories to critic fitting, predictive holdout, pilot normalization, head learning, head validation, and actor optimization before forming token batches. Head-learning and actor items need not share prefixes. All tokens in an actor trajectory are eligible for PPO, with only padding and invalid actions masked. The complete pre-action context c_t records visible text/thought prefixes, scoring text, time, remaining token allowance, and previous checkpoint potential. To state the unbiased single-trajectory estimator without a random denominator, fix a maximum thought horizon $T_{\max}$ and let m_t indicate that action t is valid. Define the finite occupation measure by

$$\mathbb{E}_{d,q}[f] := \mathbb{E}_{\tau\sim q}\left[\frac{1}{T_{\max}}\sum_{t=1}^{T_{\max}} m_t f(c_t, a_t, \xi_t)\right]. \tag{121}$$

For state-only f use $\mathbb{E}_d$. The validity of position t is known from its prefix, so conditioning on that position preserves the q continuation law. The measure need not have unit mass, but all risk, dual, and retention terms use the same fixed scale. A trajectory sum divided by $T_{\max}$ is an unbiased observation. Population normalization per valid token differs by a fixed positive factor. Dividing instead by a minibatch's random valid-token count does not have this exact finite-sample unbiasedness. The auxiliary objective uses the fixed denominator above. Actor replay can retain its standard valid-token mean because the conditional decomposition is unchanged, but aggregate bounds must specify the occupation normalization.

Compute both GAEs from the same realized trajectory length and reward schedule. The weight head reads frozen pre-action features and the sampled action, not the realized suffix, return, or GAE. The auxiliary head reads the full pre-action context, with a separate training-only path for scoring information and no sampled-action input. GAE labels, both critic encoders, reward scales,

and normalizers are detached during head learning. Optimize Eq. (75). Finite replay optimizes an empirical objective and does not supply fresh population gradients at every reuse.

After head fitting, freeze (φ, ζ) before independent validation. The base retention test checks $\widehat{C} \leq \kappa\widehat{Q}$ with the trajectory aggregation above. A failure falls back to $w = 1/2$. $\widehat{Q} = 0$ is reported as a no-reference-signal diagnostic. An empirical pass is not a population certificate. The bounded version in Appendix C.7 can enforce a population retention bound when its assumptions hold. A second selected candidate requires fresh validation rather than reuse of a rejected validation set as new evidence.

Only after the selected gate is frozen are fresh actor trajectories generated. Store each valid action, its old log probability, both GAEs, the frozen weight, the detached mixed normalized advantage, and the valid-token mask. Metadata identify policy, scorer, critic/feature, normalizer, weight, auxiliary, and validation snapshots. No weight or feature encoder is updated inside the actor PPO window. Auxiliary language-model and prediction-gate objectives retain separate backward passes. At window closure refresh the rollout law and collect new data for the changed target. Single-rollout means one thought per selected item, not one item or trajectory for the entire training process.

E.5 The prediction-mixture gate

The prediction mixture retains its base-language-model anchor and separate gate optimization. Auxiliary gate losses detach the expert logits and hidden features. A no-harm term penalizes choices that degrade prediction, and a utility-aware term can tie the mixture weight to signed thought utility. The final gate layer uses small nonzero weights: an exactly zero layer blocks upstream gradients and can leave the mixture weight constant.

Let $\beta_g(s) \in [0, 1]$ be the learned prediction-mixture gate, distinct from the learned critic-mixture weight $w_\varphi(s, a)$ and auxiliary mean function $h_\zeta(c)$. A utility-aware loss $-\mathbb{E}[\beta_g(s)\mathcal{U}(s)/\sigma_u]$ has gradient $-\mathcal{U}(s)\nabla\beta_g(s)/\sigma_u$: it favors thoughts on helpful positions and suppresses them on harmful ones. A no-harm or mixed-NLL objective can dominate this direction when most utilities are negative. A negative empirical correlation between gate weight and raw gain therefore does not by itself indicate an incorrect gradient, whereas an exactly blocked output layer prevents adaptation.

E.6 State, action, and computational cost

Let B_{text} be text batch size, P selected positions per sequence, C candidates per position, T thought length, and K potential checkpoints. The number of hidden actions is $N_{\text{act}} = B_{\text{text}}PCT$. Thought sampling scales with N_{act}, while continuation scoring scales approximately with $B_{\text{text}}PC(K+1)H$ token evaluations before caching effects. Exact prefix-value evaluation can additionally scale with the total number of partial prefixes. The principal memory objects are actor and two critic states, optimizer partitions, replayed action tensors, and temporary future-window logits. The design does not require a response-generation engine because the rollout is internal and tightly coupled to teacher-forced future evaluation. Distributed execution instead assigns separate process groups to the actor and critics. A compact single-node profile devotes most devices to the actor and one device to each critic. A larger profile uses a multi-node actor group and two independent critic groups. Replay tensors and model synchronization dominate communication, not external environment interaction.

E.7 Trajectory counts and computational cost

A rollout is one trajectory generated by the actor from its declared start to termination. Evaluating that trajectory with two critics produces paired advantages without generating a second trajectory. The distinction also applies to explicit Monte Carlo mean estimation: from M continuations $(a^{(j)}, \xi^{(j)})$ at a fixed prefix, compute

$$\widehat{\mu}_{\widetilde{A}_i}(s) = \frac{1}{M}\sum_{j=1}^{M} \widetilde{A}_i(s, a^{(j)}, \xi^{(j)}), \qquad i \in \{1, 2\}. \tag{122}$$

Both estimates use the same M generated continuations and $2M$ critic evaluations. They may be correlated, but estimating their individual means does not require separate trajectory sets. Indepen-

dent calibration and validation sets, when required by a protocol, add generation for their statistical roles rather than for the number of critics.

For the conditional-moment protocol, each selected item has $C = 1$. Denote the numbers of newly generated trajectories by n_V for critic fitting, n_{qual} for predictive holdout, n_{pilot} for normalization, n_{head} for head learning, n_{val} for candidate validation, and n_{act} for actor optimization. The two critics use the same trajectories in each set. Let n_{extra} count additional trajectories from retries, rejected attempts, or diagnostics, excluding those already counted. Then

$$\begin{aligned} n_{\mathrm{aux}} &= n_V + n_{\mathrm{qual}} + n_{\mathrm{pilot}} + n_{\mathrm{head}} + n_{\mathrm{val}} + n_{\mathrm{extra}}, \\ N_{\mathrm{roll}}^{\mathrm{mom}} &= n_{\mathrm{act}} + n_{\mathrm{aux}}. \end{aligned} \tag{123}$$

Each newly generated trajectory is counted once. Replaying stored trajectories adds optimization but no generation. The original generation is counted in the window in which it occurred. Historical reuse must still respect the snapshot and separation conditions in Appendix E.4. Trajectory counts and head-update counts are configuration parameters, not constants implied by the single-rollout identity. In particular, equally sized head-training and actor trajectory sets would contribute $2n_{\mathrm{act}}$ trajectories before validation and other trajectory sets are included.

A repeated-prefix calibration baseline with N_{anchor} anchors and M_{cal}, M_{adm}, and M_{act} continuations per anchor for calibration, admission, and actor updates generates

$$N_{\mathrm{roll}}^{\mathrm{rep}} = N_{\mathrm{anchor}}(M_{\mathrm{cal}} + M_{\mathrm{adm}} + M_{\mathrm{act}}) + n_{\mathrm{base}}^{\mathrm{rep}}, \tag{124}$$

where $n_{\mathrm{base}}^{\mathrm{rep}}$ includes prefix-bank collection, critic fitting, holdout, pilot, and additional attempts. Choosing $M_{\mathrm{cal}} = M_{\mathrm{adm}} = M$ gives $2M$ calibration/admission continuations per anchor because there are two independent phases, each sharing its trajectories between critics. The anchor-only variant contributes $N_{\mathrm{anchor}} M_{\mathrm{act}}$ actor actions despite generating their full suffixes. The moment protocol uses all valid actor-trajectory tokens, but its auxiliary trajectories do not enter the actor loss with the stated data separation. Its sampling benefit comes from sharing the learned mean function across contexts and using complete actor trajectories. Estimation error still depends on data coverage, function capacity, and optimization, so this identity alone gives no equal-accuracy sample-complexity guarantee.

For GRPO with N input items and group size G_{grp}, group-response generation contributes NG_{grp} trajectories, whose valid tokens enter actor training. At the same number of actor input items, set $n_{\mathrm{act}} = N$. With no additional GRPO generation in the compared window,

$$N_{\mathrm{roll}}^{\mathrm{mom}} < NG_{\mathrm{grp}} \quad \Longleftrightarrow \quad n_{\mathrm{aux}} < (G_{\mathrm{grp}} - 1)N. \tag{125}$$

All GRPO-based baselines in our experiments use $G_{\mathrm{grp}} = 8$, so this threshold is $n_{\mathrm{aux}} < 7N$. Additional generated data for either method must be included symmetrically. Equal input-item counts give GRPO more actor trajectories, so this inequality is not an equal actor-data comparison. At a matched number of actor trajectories, the moment protocol has additional trajectories for auxiliary training and validation, whereas GRPO distributes those actor trajectories into groups. Neither comparison establishes time or samples needed to reach a target quality.

Report generated-token count as $\sum_{\tau\ \mathrm{generated}} |\tau|$, together with the number of tokens eligible for actor updates. Repeated-prefix suffixes and full item trajectories can have different lengths, so trajectory counts translate directly to token savings only with comparable average lengths. Total computation also includes reward scoring, actor updates, both critics' inference and training, feature construction, and weight/mean-head optimization. Once frozen features and GAE labels are cached, further head passes require no new rollout, but increase computation and optimize the same empirical sample. Head parameters, optimizer states, and cached features add memory, and constructing the mean head's scoring-context features also costs computation. Removing prefix replication changes sampling, but speedups depend on total computation and implementation.

E.8 Refreshing and recovering a training window

The trainer retains value-head warmup, full-critic warmup, alternating training, and actor-paused recovery. Backbone unfreezing resets predictive qualification. Predictive failure pauses actor updates and triggers return-data refresh and critic refitting. After qualification, freeze critics and normalizers, train the heads, and validate a frozen candidate's signal retention. Failed retention returns to the feasible fixed mean. It does not certify a smaller drift. The base protocol has no hard actor-release test

based only on a small saddle loss. If a drift-certified mode is reported, it must additionally provide all error bounds in Corollary 1. Changing rollout, scorer, critics, features, or normalizer changes the moment target and requires renewed fitting and independent evaluation. A whole-window drift claim also needs a common actor sensitivity bound B.

AI USE STATEMENT

Generative AI tools assisted with manuscript editing, exploration of the conditional-calibration design, and drafting of its analysis. The authors remain responsible for independently verifying the technical content, claims, citations, and conclusions before submission.