ENVIRONMENT-STALENESS-AWARE REPLAY FOR LANGUAGE AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reusing past interactions can make reinforcement learning for large language model agents more efficient, but changes in tool reliability can render those interactions misleading even when the policy remains unchanged. We introduce Environment-Staleness-Aware Replay (ESR) to retain useful historical experience while accounting for changes in the environments that produced it. ESR executes the same actions from restored states in both environment versions and uses a regularized least-squares fit to estimate how the probabilities of tool outcomes have changed. It combines these estimates with policy corrections to reweight complete interaction sequences, excludes replay without sufficient execution evidence, and counts probing and restoration toward the interaction budget. On AppWorld, ESR improves mean post-change task-completion area under the learning curve by 2.44–4.37 points on a 0–100 scale over training with fresh interactions alone across three models and two test splits, with all methods sharing the same total interaction budget. These findings identify environment staleness as a distinct source of replay error and support evidence-based experience reuse in restorable environments with changing tool reliability.

1 INTRODUCTION

Large language model (LLM) agents solve multi-step digital tasks through reasoning and environmental actions (Yao et al., 2023). Reinforcement learning (RL) improves these capabilities through interaction and long-horizon credit assignment (Feng et al., 2025). Training across many tools and persistent states (Trivedi et al., 2024) requires costly repeated generation and execution, motivating reuse of past interactions.

Experience replay reduces generation cost while balancing policy staleness and sample diversity (Arnal et al., 2026). Prioritized replay favors informative experience (Schaul et al., 2016), while FreshPER discounts aging priorities as policies evolve (Ma et al., 2026). Yet past interactions also depend on the environment that produced them, whose changes can alter their learning value.

When a tool becomes less reliable, old successful trajectories overrepresent outcomes that are now harder to obtain, even under an unchanged policy. Repeated calls amplify this discrepancy: trajectories remain *on-policy* but become *off-environment*. Policy ratios correct action selection; changed outcome probabilities require an environment ratio. Equally old trajectories can be affected differently, motivating selective reuse to preserve interaction savings.

Prior RL work corrects source–target dynamics mismatch through reward augmentation: DARC addresses source-domain learning (Eysenbach et al., 2021), and DARA extends this approach to offline transfer (Liu et al., 2022). For multi-turn LLM agents, we instead correct historical trajectories for the current task objective under a shared budget for correction and fresh interaction. The benefit of correcting replay must therefore justify the fresh interactions displaced by collecting evidence of environment change.

We introduce *Environment-Staleness-Aware Replay* (ESR) for restorable environments with changing tool reliability, fixed interfaces and outcome-conditioned dynamics, and overlapping source and current support (Figure 1). Paired executions of the same restored state and action across versions support a regularized least-squares estimate of changed tool-outcome likelihoods. ESR combines

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

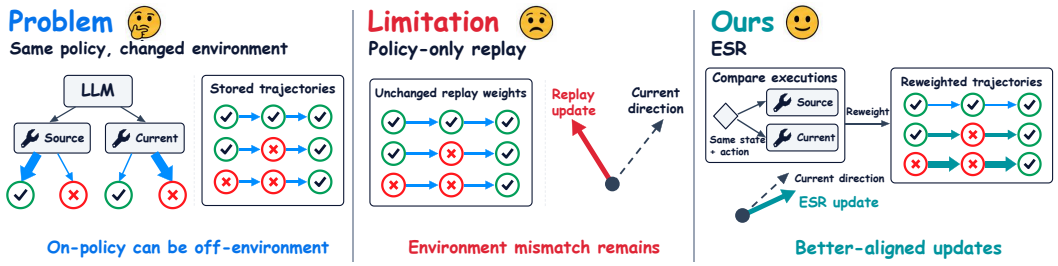

Figure 1: **On-policy experience can be off-environment. Left:** Tool reliability changes while the policy stays fixed. **Middle:** Policy-only replay leaves this environment mismatch unresolved. **Right:** ESR compares source and current executions to reweight historical trajectories for the current environment. Recorded outcomes stay unchanged, and failure-containing trajectories can gain weight. Line widths and update directions are schematic.

this evidence with policy correction over complete trajectories, restricts reuse when evidence is insufficient, and charges probing, restoration, and fresh interaction to one budget.

We make three contributions:

- **We reveal environment staleness in LLM-agent replay.** Frozen-policy diagnostics and a formal counterexample show that tool reliability changes can reverse replay gradients even without policy drift.
- **We propose ESR for reuse across environment versions.** The method turns limited execution evidence into a trajectory-level correction without an auxiliary neural network. Our theory establishes when exact correction recovers the current-environment policy gradient and identifies limits imposed by missing support and incomplete information.
- **ESR improves adaptation under a shared interaction budget.** On AppWorld, ESR improves mean post-switch task-completion AUC by 2.44–4.37 points on a 0–100 scale over fresh-only training across three backbones and two test splits, including probe and restoration costs. Controlled ALFWorld and WebShop evaluations show consistent gains.

2 Related Work

Multi-turn agent learning. Multi-turn RL addresses sparse feedback through hierarchical optimization and finer credit assignment (Zhou et al., 2024; Feng et al., 2025). Subsequent work expands training experience through learned simulators (Chen et al., 2026) and feedback-driven task or environment adaptation (Wang et al., 2026; Yang et al., 2026). ESR complements these approaches by correcting historical transitions after tool reliability changes, with a fixed task distribution and reward.

Experience replay and policy staleness. Prioritized replay selects informative transitions (Schaul et al., 2016), while recent LLM studies formalize the trade-off between generation cost, diversity, and stale-policy data (Arnal et al., 2026). FreshPER discounts priorities by age (Ma et al., 2026); advantage-prioritized GRPO replay combines age eviction with fresh rollouts (Yoo et al., 2026). ESR estimates changes in tool-response probabilities, addressing replay mismatch even when the behavior and current policies coincide.

Off-dynamics learning and ratio estimation. DARC uses classifier-estimated dynamics ratios to modify rewards (Eysenbach et al., 2021), DARA develops reward augmentation for offline dynamics adaptation (Liu et al., 2022), and H2O uses dynamics ratios to weight Bellman errors alongside value regularization (Niu et al., 2022). ESR instead corrects complete-trajectory policy gradients using paired executions and least-squares ratio estimation (Kanamori et al., 2009). It combines tool–outcome ratios fitted at restored anchors with policy factors across turns, charging probes to the interaction budget. This requires executable source versions, complete-state restoration, overlapping support, and reliability-mixture changes.

3 PRELIMINARY STUDY

Can replay become stale even when the policy is not? We freeze the agent policy and vary only the execution reliability of eight tools, comparing historical replay gradients with fresh gradients from the current environment. Across 56 drift–call-count settings and 24 seeds per setting, we measure gradient cosine alignment A_{grad}; negative values indicate directional reversal. Here δ is the mean absolute change in tool reliability, and n_{tool} counts affected calls per trajectory. Appendix B.1 gives the setup and metric definitions.

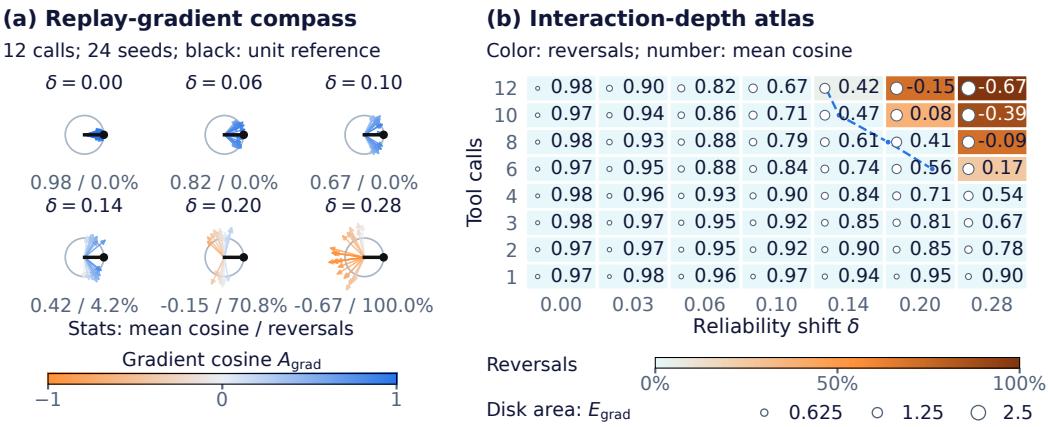

Figure 2: **On-policy replay can become off-environment. (a)** Replay gradients relative to the fresh reference for twelve affected calls (24 seeds per setting). **(b)** Reversal fraction (color), mean gradient cosine (number), and mean relative gradient error (disk area) across drift and call count. The dashed line marks the interpolated mean-alignment threshold of 0.5.

Reliability drift can reverse the replay direction. With twelve affected calls, mean alignment falls from 0.98 at $\delta = 0$ to -0.67 at $\delta = 0.28$, where all 24 seeds show reversals (Figure 2a). Environment drift can therefore redirect replay gradients even under a fixed policy.

The mismatch grows with interaction depth. At $\delta = 0.28$, four affected calls retain positive mean alignment (0.54) with no reversals, whereas eight calls yield -0.09 alignment and 17 of 24 reversals (Figure 2b). The decline from four to eight to twelve calls shows how repeated exposure amplifies the mismatch.

The policy likelihood ratio is identically one in this study, so policy-only correction leaves the replay weights unchanged. We next develop an environment correction that combines evidence across a trajectory.

4 METHOD

ESR probes historical state–action pairs in source and current environments, fits a centered least-squares correction to tool–outcome frequencies, and combines it with policy likelihood ratios to weight replayed trajectories (Figure 3). Each source version and task fold requires one regularized covariance solve. Evidence checks restrict replay to supported tools under a fixed probe budget.

4.1 REPLAY UNDER CHANGES IN TOOL RELIABILITY

Let x be a task, s_t the complete environment state, o_t an observation, and $h_t = (x, o_0, a_0, \ldots, o_t)$ the observable history. Action a_t is a complete sampled generation turn invoking at most one affected tool. A trajectory $\tau = (x, s_0, o_0, a_0, s_1, o_1, \ldots, a_{T-1}, s_T, o_T)$ has $T \le H$ turns and terminal reward $R(\tau) \in [0, 1]$. Its law in version v and our objective are

$$p_{\theta,v}(\tau) = q(x, s_0, o_0) \prod_{t<T} \pi_\theta(a_t \mid h_t) P_v(s_{t+1}, o_{t+1} \mid s_t, a_t),$$
$$J_\star(\theta) = \mathbb{E}_{p_{\theta,v_\star}}[R(\tau)], \tag{1}$$

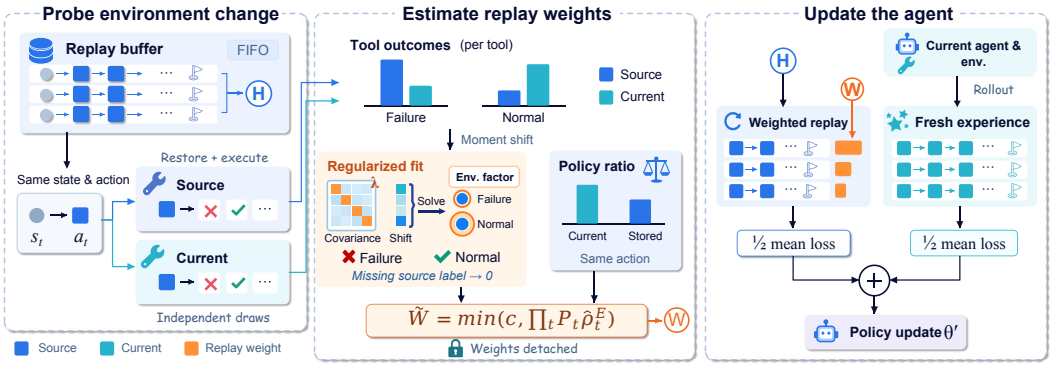

Figure 3: **Overview of ESR.** Restored state–action pairs are executed independently in source and current environments. A regularized covariance solve estimates environment factors, multiplied with policy ratios over all turns. The product is capped once and fixed during differentiation; missing either source outcome gives zero replay weight. A nonempty buffer gives equal coefficients to fresh and weighted replay mean losses. Blue: source trajectories; cyan: current experience; orange: replay weights. Bar heights and halo sizes are schematic.

Here q is the shared initial law and $v_\star$ the current version, fixed within each update round. Versions share state spaces, action interfaces, stopping rules, and rewards. Training requires executable source versions and restoration of all transition-relevant state, including interpreter and session state; policies observe only h_t.

Tool reliability model. Write $u = (s, a)$ and $z = (s', o')$. Each affected tool has a fixed identifier $k \in \{1, \dots, K\}$ and returns label $y \in \{0, 1\}$ for wrapper failure or normal execution, respectively. Reliability changes follow

$$P_v(z \mid u) = \sum_{y=0}^{1} p_v(y \mid k(u)) G_y(z \mid u), \qquad \rho_v^E(u, z) = \frac{p_{v_\star}(y(z) \mid k(u))}{p_v(y(z) \mid k(u))}. \tag{2}$$

The invariant kernels G_y have disjoint, label-identified supports. Only tool-specific outcome probabilities change; both source probabilities must be positive for every affected replayed tool. Transitions without affected calls have ratio one. Probes can therefore be pooled across restored anchors of each tool. Within-label changes require a richer model; replacing one deterministic response with another creates missing support and requires fresh trajectories. Appendix A.1 specifies execution and record requirements.

Correcting both sources of staleness. Record i was collected by behavior policy β_i in version v_i. At frozen parameters θ_n, its exact trajectory ratio is

$$W_i = \prod_{t<T_i} \frac{\pi_{\theta_n}(a_{i,t} \mid h_{i,t})}{\beta_i(a_{i,t} \mid h_{i,t})} \rho_{v_i}^E(u_{i,t}, z_{i,t}). \tag{3}$$

The behavior policy must cover current-policy actions, ensuring $p_{\theta_n, v_\star} \ll p_{\beta_i, v_i}$ under Equation (2). Policy and environment ratios correct action selection and transitions, respectively; both span the full trajectory because terminal rewards depend on the entire interaction.

4.2 Paired least-squares correction

Paired sampling and fixed features. Split training tasks into two fixed folds. For trajectories in fold f, draw source anchors only from fold $1 - f$, independently of stored rewards and outcomes. Restore isolated source/current copies at each anchor and execute with independent randomness: $z^- \sim P_v(\cdot \mid u)$, $z^+ \sim P_{v_\star}(\cdot \mid u)$. The anchor–outcome laws $Q_{v,f}^- = \nu_{v,f} P_v$ and $Q_{v,f}^+ = \nu_{v,f} P_{v_\star}$ share an anchor law that cancels, yielding $dQ_{v,f}^+ / dQ_{v,f}^- = \rho_v^E$ without an anchor-visitation factor.

Encode tool–outcome pairs as unscaled one-hot vectors $\phi(u, z) = e_{2(k(u)-1)+y(z)+1} \in \mathbb{R}^d$, with $d = 2K$ and standard basis e_j. Exclude invalid identifiers and missing labels. Fixing a source version

and scored fold, write $\phi_j^{\pm} = \phi(u_j, z_j^{\pm})$ for the m valid pairs and compute

$$\widehat{\mu} = \frac{1}{m}\sum_j \phi_j^-, \qquad \widehat{\Delta} = \frac{1}{m}\sum_j (\phi_j^+ - \phi_j^-), \qquad \widehat{C} = \frac{1}{m}\sum_j (\phi_j^- - \widehat{\mu})(\phi_j^- - \widehat{\mu})^\top. \tag{4}$$

Thus $\widehat{\mu}$, $\widehat{\Delta}$, and $\widehat{C}$ are the source frequencies, paired frequency shift, and source covariance. Appendix A.1.2 specifies the tool manifest and label rules.

Least-squares fitting around unit weight. Write the raw ratio as $r_w(u,z) = 1 + (\phi(u,z) - \widehat{\mu})^\top w$. Centering keeps its empirical source mean at one for any w. Least-squares density-ratio fitting (Kanamori et al., 2009) gives

$$\widehat{w} = \arg\min_{w \in \mathbb{R}^{2K}} \left\{ \frac{1}{2} w^\top \widehat{C} w - \widehat{\Delta}^\top w + \frac{\lambda}{2}\|w\|_2^2 \right\}, \qquad \lambda > 0. \tag{5}$$

The terms penalize source-probe deviations from unit weight, fit the frequency shift, and regularize poorly sampled directions. Differentiation yields

$$(\widehat{C} + \lambda I_{2K})\widehat{w} = \widehat{\Delta}, \qquad \widehat{r}(u,z) = 1 + (\phi(u,z) - \widehat{\mu})^\top \widehat{w}. \tag{6}$$

For one-hot features, $\widehat{C} = \mathrm{diag}(\widehat{\mu}) - \widehat{\mu}\widehat{\mu}^\top$. This diagonal-minus-rank-one structure gives an $O(K)$ fit with strictly positive raw ratios (Appendix A.2). Without ridge regularization, the population fit matches the target moments while minimizing deviation from unit weight. Under Equation (2), it equals the true ratio on probe support; Appendix D.2 proves this projection result.

Evidence required for a replay factor. For changed tool k, let $N^-_{v,f,k,y}$ count valid source-probe outcomes labeled y in fold f's fit. Set $\widehat{\rho}^E_{v,f} = \widehat{r}$ only when $N^-_{v,f,k,0} > 0$, $N^-_{v,f,k,1} > 0$, the record is valid, and the fit is finite; otherwise use zero. Current-version transitions and declared unchanged tools have factor one; unprobed tools receive zero regardless of numerical solvability. State-dependent changes require coverage conditions beyond this tool-specific reliability model.

A fixed probe budget. Let $N_{\max}$ be the total call budget and N_{call} the calls already spent. Let L_F, L_P bound the calls for one fresh trajectory and one probe pair, including restoration. Reserve $N_F = BL_F$ for a fresh batch of size B. With probe fraction α and $N_{\max} - N_{\mathrm{call}} \geq N_F$, the probe-attempt quota is

$$m_n = \min\left\{ \left\lfloor \frac{\alpha N_F}{(1-\alpha)L_P} \right\rfloor, \left\lfloor \frac{N_{\max} - N_{\mathrm{call}} - N_F}{L_P} \right\rfloor \right\}, \qquad 0 \leq \alpha < 1. \tag{7}$$

The bounds limit the probe share of reserved calls and enforce the remaining budget. Attempts cycle evenly through eligible source-version, scored-fold, and tool triples, sampling anchors uniformly within each; a persistent cursor continues across rounds. Failed pairs consume attempts and calls but add no moments. No eligible triple means no probes; version changes clear fits and counts. Appendix A.3 specifies eligibility, ordering, and stopping.

4.3 One trajectory-weighted update

Collect B fresh trajectories $\mathcal{F}_n$ under $(\pi_{\theta_n}, v_\star)$ and sample B old trajectories $\mathcal{H}_n$ uniformly with replacement from the nonempty FIFO buffer. Combine environment factors with policy ratios using the stored behavior probabilities:

$$\begin{aligned} \widehat{\ell}_i &= \sum_{t<T_i} \left[\log \pi_{\theta_n}(a_{i,t} \mid h_{i,t}) - \log \beta_i(a_{i,t} \mid h_{i,t}) + \log \widehat{\rho}^E_{i,t} \right], \\ \widetilde{W}_i &= \exp\Big(\min\{\log c, \widehat{\ell}_i\} \Big), \qquad c \geq 1. \end{aligned} \tag{8}$$

The cap c applies once to the full trajectory. With $\log 0 = -\infty$, any ineligible transition sets its trajectory weight to zero. Records with missing or invalid likelihood, reward, or transition metadata are skipped during score evaluation and contribute zero. Log probabilities include every sampled action token under the actual decoding distribution.

Define $S_i(\theta) = \sum_{t<T_i} \log \pi_\theta(a_{i,t} \mid h_{i,t})$ and let b_n be the previous fresh batch's mean reward, initialized to zero. The update minimizes

$$\mathcal{L}_n(\theta) = -\frac{1}{2B}\left[\sum_{i \in \mathcal{F}_n} \mathrm{sg}(R_i - b_n) S_i(\theta) + \sum_{i \in \mathcal{H}_n} \mathrm{sg}\Big(\widetilde{W}_i (R_i - b_n)\Big) S_i(\theta) \right], \tag{9}$$

where sg stops gradients through rewards and weights. Take one step $\theta_{n+1} = \theta_n - \epsilon_n \nabla \mathcal{L}_n(\theta_n)$ with step size ϵ_n, then append the fresh batch to the buffer. For a nonempty buffer, the fresh mean keeps coefficient $1/2$ even if all replay weights are zero. With an empty buffer, use only the fresh mean loss, with coefficient one. Rejected old records still count toward the denominator; losses use unnormalized complete-trajectory scores.

Appendix A.3.2 gives the training procedure. The analysis below establishes exact gradient recovery, and Appendix D.5 bounds the bias from finite probes, ridge, evidence exclusions, and capping, including the accumulation of local ratio errors across turns.

4.4 Theoretical justification

Under the reliability model in Equation (2), environment correction recovers the current gradient; policy correction alone recovers the source-environment gradient. The result assumes shared initial laws, rewards, and stopping rules, together with the finite-horizon, support, paired-probe validity, and score-regularity conditions in Appendix D.1.

Theorem 1 (Identifiable correction and current-environment gradients)**.** *Under these assumptions, the unregularized population one-hot fit recovers $\rho_v^E = p_{v_\star}(y \mid k)/p_v(y \mid k)$ at every anchor of a covered tool. Fix θ, β, v, and $b \in [0, 1]$ independently of the source trajectories. Write $P_0 = p_{\beta,v}$, $A = \prod_{t<T} \pi_\theta(a_t \mid h_t)/\beta(a_t \mid h_t)$, $L = \prod_{t<T} \rho_v^E(u_t, z_t)$, and $\mathsf{s}_\theta = \sum_{t<T} \nabla \log \pi_\theta(a_t \mid h_t)$, with zero score and contribution when $A = 0$. Then*

$$\begin{aligned} \mathbb{E}_{P_0}[AL(R-b)\mathsf{s}_\theta] &= \nabla J_\star(\theta), \\ \mathbb{E}_{P_0}[A(R-b)\mathsf{s}_\theta] &= \nabla J_v(\theta), \qquad J_v(\theta) = \mathbb{E}_{p_{\theta,v}} R. \end{aligned} \tag{10}$$

For independent fixed-source trajectories, the corrected empirical mean converges almost surely to $\nabla J_\star(\theta)$.

Proof sketch. The shared probe-anchor law cancels, and the one-hot features represent the environment ratio exactly. Multiplying policy and environment ratios gives $dp_{\theta,v_\star}/dP_0 = AL$; using A alone changes only the policy. The zero-mean score identity yields both gradients, and target integrability gives almost-sure convergence. Appendix D.4 proves the result using the identification theorem in Appendix D.3. □

Even with $\beta = \pi_\theta$, swapping two tools' reliabilities can reverse the gradient: $\nabla J_\star = -\nabla J_v \neq 0$ (Proposition 1). Appendix D gives sharp bounds on the bias from omitting environment correction and the limits imposed by discarded features and missing support. Appendix D.5 analyzes practical correction error and the dependence introduced by an adaptive buffer.

5 Experiments

We evaluate budget-adjusted replay gains (Q1; Section 5.2), generalization across tasks and held-out domains (Q2; Section 5.3), the roles of individual components and gradient correction (Q3; Section 5.4), and when correction is unnecessary or insufficient (Q4; Section 5.5).

5.1 Experimental setup

Benchmarks and metrics. We use AppWorld (Trivedi et al., 2024) (TGC/SGC), ALFWorld (Shridhar et al., 2021) (success rate), and WebShop (Yao et al., 2022) (score/success rate). Adaptation is measured by success-rate AUC against charged environment calls; AppWorld-to-BFCL V3 transfer is evaluated separately (Appendix B.2).

Comparison protocol. Controlled runs match backbones, tasks, interfaces, and charged interaction budgets across three seeds. Count-ratio isolates estimator choice (Appendix B.4). Native panels retain source evaluation protocols and method-specific training resources.

Backbones and implementation. We use Qwen2.5-7B-Instruct, Qwen3-4B, and Llama-3.1-8B-Instruct (configuration: Appendix B.6). Task scores use a 0–100 scale; higher is better. Bold/underline mark the highest/second-highest means within each backbone/protocol panel; "–" denotes an unreported source metric.

5.2 Q1: Does replay help after paying for correction?

Controlled adaptation. Under the shared 10^6-call budget, ESR has the highest mean AUC in all six AppWorld backbone–split combinations, exceeding fresh-only by 2.44–4.37 points; policy-only replay remains below fresh-only throughout (Table 1). ESR also leads on ALFWorld/WebShop, with 5.01/4.70-point AUC gains over fresh-only (Table 6).

Table 1: **AppWorld adaptation.** Post-switch TGC AUC ($\times 100$), three seeds. Budget: 10^6 charged calls; normal execution: $0.7 \rightarrow 0.5$ at 40%. N/C: Test-N/Test-C. Subscripts: s.d.; bold/underline: highest/second-highest mean.

	Qwen2.5-7B		Qwen3-4B		Llama-3.1-8B	
Method	N	C	N	C	N	C
Fresh-only	$21.84_{\pm 0.93}$	$7.92_{\pm 0.61}$	$29.86_{\pm 1.02}$	$13.41_{\pm 0.77}$	$25.68_{\pm 1.11}$	$10.49_{\pm 0.74}$
Buffer-flush	$22.63_{\pm 0.88}$	$8.35_{\pm 0.55}$	$30.74_{\pm 0.91}$	$13.96_{\pm 0.71}$	$26.39_{\pm 1.05}$	$10.97_{\pm 0.69}$
Uniform replay (Arnal et al., 2026)	$18.71_{\pm 1.27}$	$6.44_{\pm 0.72}$	$25.18_{\pm 1.45}$	$10.84_{\pm 0.96}$	$21.91_{\pm 1.33}$	$8.46_{\pm 0.85}$
Policy-only replay	$19.16_{\pm 1.12}$	$6.58_{\pm 0.69}$	$25.63_{\pm 1.26}$	$11.17_{\pm 0.83}$	$22.34_{\pm 1.19}$	$8.69_{\pm 0.78}$
PER (Schaul et al., 2016)	$18.96_{\pm 1.36}$	$6.31_{\pm 0.81}$	$25.41_{\pm 1.39}$	$10.96_{\pm 0.88}$	$22.12_{\pm 1.42}$	$8.53_{\pm 0.81}$
FreshPER (Ma et al., 2026)	$21.47_{\pm 1.01}$	$7.64_{\pm 0.63}$	$28.09_{\pm 1.09}$	$12.71_{\pm 0.76}$	$24.71_{\pm 1.08}$	$9.91_{\pm 0.72}$
FreshPER + sampling IS (Ma et al., 2026)	$21.31_{\pm 0.94}$	$7.78_{\pm 0.58}$	$28.37_{\pm 1.01}$	$12.88_{\pm 0.73}$	$24.92_{\pm 1.02}$	$10.08_{\pm 0.68}$
DARC-style (Eysenbach et al., 2021)	$23.71_{\pm 0.86}$	$8.92_{\pm 0.57}$	$31.56_{\pm 0.94}$	$14.76_{\pm 0.66}$	$27.06_{\pm 0.93}$	$11.69_{\pm 0.62}$
Count-ratio correction	$\underline{25.54}_{\pm 0.72}$	$\underline{9.87}_{\pm 0.49}$	$\underline{33.26}_{\pm 0.76}$	$\underline{15.82}_{\pm 0.58}$	$\underline{28.61}_{\pm 0.79}$	$\underline{12.51}_{\pm 0.55}$
ESR (ours)	$\mathbf{26.21}_{\pm 0.68}$	$\mathbf{10.36}_{\pm 0.46}$	$\mathbf{34.02}_{\pm 0.70}$	$\mathbf{16.43}_{\pm 0.54}$	$\mathbf{29.34}_{\pm 0.74}$	$\mathbf{13.02}_{\pm 0.51}$

Estimator choice and cost. Count-ratio changes only the estimator, using smoothed counts (Appendix B.4). Under Eq. (2), both target the same ratios (Appendix D.3). ESR exceeds count-ratio by 0.49–0.76 mean AUC points. On Qwen2.5-7B, paired 95% intervals for ESR's gains include zero against count-ratio but are positive against fresh-only and policy-only (Table 12). Full ESR uses 109.6K probe/restoration calls and 72.4 wall-hours versus 62.3 for fresh-only (Appendix C.5).

Takeaway (Q1). Correcting for current transitions can improve on fresh-only training under a fixed call budget, with additional wall-clock cost.

5.3 Q2: Do the gains extend across tasks and domains?

Native task performance. Under native protocols, ESR improves AppWorld TGC over CoEvolve on both backbones and splits (Table 2). WebShop success reaches 76.7% versus GiGPO's 75.2%; ALFWorld unseen success reaches 65.8% versus RLAnything's 63.6% (Table 3). DreamGym comparisons and detailed gains are in Appendix C.3.

Table 2: **AppWorld task completion.** Baselines: CoEvolve (Yang et al., 2026). Greedy decoding, 30 turns, 4,096 response tokens; no injected failures. Training resources are method-specific.

		Test-N		Test-C	
Backbone	Method	TGC	SGC	TGC	SGC
Qwen2.5-7B-Instruct	Base model (Yang et al., 2026)	1.19	0.00	0.72	0.00
	GRPO (Shao et al., 2024)	26.78	–	–	–
	CoEvolve (Yang et al., 2026)	<u>27.98</u>	<u>12.50</u>	<u>8.39</u>	<u>2.16</u>
	ESR (ours)	**30.36**	**14.29**	**10.31**	**2.88**
Qwen3-4B	Base model (Yang et al., 2026)	16.67	5.36	7.91	2.16
	GRPO (Shao et al., 2024)	28.57	–	–	–
	CoEvolve (Yang et al., 2026)	<u>35.71</u>	<u>14.28</u>	<u>17.03</u>	<u>6.47</u>
	ESR (ours)	**38.10**	**16.07**	**19.18**	**7.91**

Held-out transfer. AppWorld-trained Qwen3-4B reaches 47.50% on BFCL V3 Multi-Turn Base versus CoEvolve's 45.00%, without BFCL training or probes (Table 8).

Takeaway (Q2). Environment-specific correction is compatible with reusable tool-use skills, as shown by transfer without target-domain probes.

Table 3: **Native agent performance.** Qwen2.5-7B-Instruct. Baselines: GiGPO (Feng et al., 2025) (A), RLAnything (Wang et al., 2026) (B). Subscripts: s.d.; ESR uses three seeds. Training resources are method-specific.

A. WebShop *GiGPO protocol: temperature* 0.4*, 15 turns*		
Method	Score	Success rate
ReAct (Yao et al., 2023)	46.2	19.5
PPO (Schulman et al., 2017)	$81.4_{\pm 3.1}$	$68.7_{\pm 5.1}$
RLOO (Ahmadian et al., 2024)	$80.3_{\pm 3.2}$	$65.7_{\pm 4.0}$
GRPO (Shao et al., 2024)	$79.3_{\pm 2.8}$	$66.1_{\pm 3.7}$
GiGPO, with std. (Feng et al., 2025)	$84.4_{\pm 2.9}$	$72.8_{\pm 3.2}$
GiGPO, without std. (Feng et al., 2025)	$\underline{86.2}_{\pm 2.6}$	$\underline{75.2}_{\pm 3.8}$
ESR (ours)	$\mathbf{87.3}_{\pm 2.2}$	$\mathbf{76.7}_{\pm 2.9}$
B. ALFWorld *RLAnything protocol: temperature* 0.8*, 60 turns*		
Method	Seen (140)	Unseen (134)
Base model (Wang et al., 2026)	39.0	44.9
RLAnything: policy only (Wang et al., 2026)	51.1	59.3
RLAnything: policy + reward (Wang et al., 2026)	55.6	61.8
RLAnything (Wang et al., 2026)	$\underline{60.2}$	$\underline{63.6}$
ESR (ours)	$\mathbf{62.5}_{\pm 1.2}$	$\mathbf{65.8}_{\pm 1.0}$

5.4 Q3: What makes corrected replay useful?

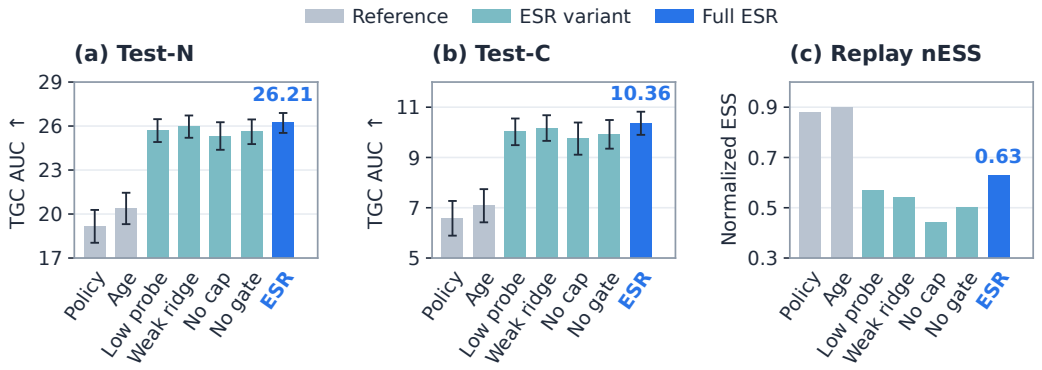

Figure 4: **Component effects on adaptation and replay weights.** Selected rows of Table 9: Qwen2.5-7B/AppWorld, $0.7 \rightarrow 0.5$ at 40% of 10^6 charged calls. Bars show means; error bars in (a,b) are sample s.d. across three seeds. Age: $\tau = 500$; low probe: $\alpha = 0.05$; weak ridge: $\lambda = 10^{-4}$. Panel (c) shows nESS without uncertainty bars. The y-axes start at 17, 5, and 0.30, respectively.

Component ablations. Age-only weighting reaches 0.90 nESS yet trails ESR by 5.83/3.28 Test-N/Test-C AUC points (Figure 4). Removing the cap lowers nESS from 0.63 to 0.44 and AUC by 0.89/0.61 points. Probe, ridge, and gate ablations further link replay quality to evidence coverage and weight concentration (Appendix C.5).

Frozen-policy diagnostic. With policy ratios fixed at one, ESR raises gradient cosine from 0.84/0.31 to 0.982/0.931 for short/long trajectories, isolating accumulated environment mismatch (Table 10).

Training-time diagnostic. Figure 5 compares replay gradients with an independent fresh-gradient estimate across three Qwen2.5-7B/AppWorld runs. It plots cosine d_m and $\log_2$ norm ratio s_m, with reference $(1, 0)$ (definitions and protocol: Appendix B.8). Immediately after the switch, ESR raises mean cosine from 0.26 to 0.89 and reduces median $|s_m|$ from 0.36 to 0.06. Across 1,056 matched batches, direction improves in 90.4% and both direction and absolute scale mismatch improve in 75.0%. The gap narrows as stale records are replaced (Appendix C.6).

Takeaway (Q3). High nESS can coexist with poor gradient alignment. Fresh-gradient comparisons reveal the directional benefit of correction, especially just after a version change.

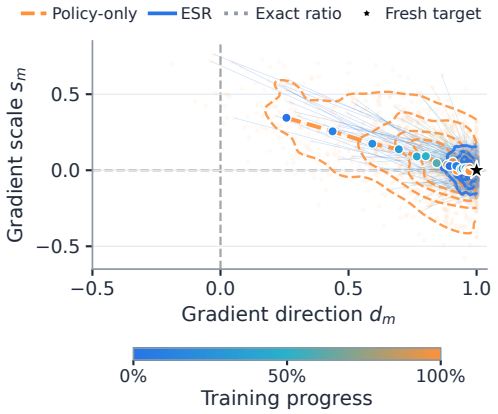

Figure 5: **Training-time gradient phase portrait.** Dots show 3,168 records from 1,056 matched batches; contours distinguish the three methods. Arrows connect 132 policy-only/ESR pairs, and curves join checkpoint means. Arrow/node color indicates training progress; the star marks the fresh-gradient target $(1, 0)$. Appendix B.8 details the sampling.

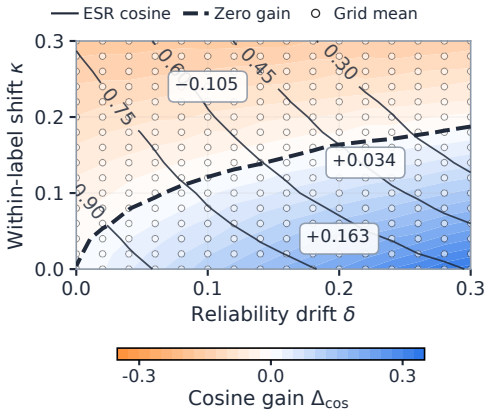

Figure 6: **Model-misspecification landscape.** A score-level illustration with 5,376 paired records over 256 grid cells. Color shows mean cosine gain $\Delta_{\cos}$ across 21 records per cell: blue indicates positive gain and orange negative gain. Solid contours show ESR cosine; the dashed contour interpolates zero gain. Dots mark grid locations.

5.5 Q4: When is correction unnecessary or insufficient?

No-change controls. On Qwen2.5-7B/AppWorld, static ESR exactly matches policy-only replay. A null switch incurs probe costs with unchanged transitions; its Test-N/Test-C differences are $-0.35/-0.22$ points, with one-sided 95% lower bounds $-0.73/-0.55$ above the -1-point non-inferiority margin (Table 11).

Model misspecification. ESR assumes invariant within-label response kernels (Eq. (2)). Figure 6 uses a stochastic score construction to illustrate sensitivity to reliability drift δ and within-label changes κ. Increasing κ narrows the region with positive cosine gain $\Delta_{\cos} = d_{\mathrm{ESR}} - d_{\mathrm{PO}}$; at $(\delta, \kappa) = (0.24, 0.24)$, mean gain is -0.086. The model and numerical details are in Appendix B.10; Proposition 2 establishes the corresponding identifiability limitation.

Practical scope. The landscape is a score-level illustration of the identifiability limit, not a training-performance evaluation. More probes can reduce sampling uncertainty for represented reliability changes, but cannot resolve changes hidden within an outcome class. The evidence gate excludes tools lacking the required source-outcome observations; fresh interaction remains necessary for outcomes outside historical support. These conditions delimit the setting in which the shared-budget gains in Q1 apply, alongside the reported wall-clock overhead.

Takeaway (Q4). Drift within an indistinguishable feature cell requires richer observations or fresh experience; more probes of the same labels leave the ambiguity unresolved.

6 Conclusion

We showed that changes in tool reliability can misdirect replay updates even when the agent's policy is unchanged. ESR addresses this mismatch by turning paired execution evidence into a lightweight correction of historical trajectories. Experiments on AppWorld, ALFWorld, and WebShop demonstrate that accounting for the execution environment improves adaptation under matched interaction budgets that include probing and restoration costs. The theory and experiments identify environment compatibility as a distinct factor in replay and show why reuse should be evaluated after accounting for evidence-collection costs. Extending correction beyond reliability shifts will require features that capture changes within outcome classes, with fresh interaction retained wherever historical trajectories lack support. Jointly optimizing probe allocation and experience reuse could extend these benefits as tool behavior evolves over longer training horizons.

AI USE STATEMENT

We used generative AI tools to assist with checking theoretical proofs, searching for relevant literature, interpreting results, polishing the manuscript's language, and preparing scientific figures. The authors remain responsible for the accuracy, originality, and integrity of the entire submission, including the correctness of the theoretical arguments, the accuracy of citations, and the faithful presentation of results in the text and figures.

REFERENCES

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting REINFORCE-style optimization for learning from human feedback in LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12248–12267. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.662. URL https://aclanthology.org/2024.acl-long.662/.

Charles Arnal, Vivien Cabannes, Taco Cohen, Julia Kempe, and Rémi Munos. Efficient RL training for LLMs with experience replay, 2026. URL https://arxiv.org/abs/2604.08706.

Zhaorun Chen, Zhuokai Zhao, Kai Zhang, Bo Liu, Qi Qi, Yifan Wu, Tarun Kalluri, Xuefei Cao, Yuanhao Xiong, Haibo Tong, Huaxiu Yao, Hengduo Li, Jiacheng Zhu, Xian Li, Dawn Song, Bo Li, Jason E. Weston, and Dat Huynh. Scaling agent learning via experience synthesis. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/c53793534833d30a40f3d18352735519-Abstract-Conference.html.

Benjamin Eysenbach, Shreyas Chaudhari, Swapnil Asawa, Sergey Levine, and Ruslan Salakhutdinov. Off-dynamics reinforcement learning: Training for transfer with domain classifiers. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/pdf/2006.13916.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for LLM agent training. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1544. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/420c9f777c0b4f78d515e53cf74d58b2-Abstract-Conference.html.

Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10(48):1391–1445, 2009. URL https://jmlr.org/papers/v10/kanamori09a.html.

Jinxin Liu, Hongyin Zhang, and Donglin Wang. DARA: Dynamics-aware reward augmentation in offline reinforcement learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=9SDQB3b68K.

Weiyu Ma, Yongcheng Zeng, Yan Song, Xinyu Cui, Jian Zhao, Xuhui Liu, and Mohamed Elhoseiny. Freshness-aware prioritized experience replay for LLM/VLM reinforcement learning, 2026. URL https://arxiv.org/abs/2604.16918.

Haoyi Niu, Shubham Sharma, Yiwen Qiu, Ming Li, Guyue Zhou, Jianming Hu, and Xianyuan Zhan. When to trust your simulator: Dynamics-aware hybrid offline-and-online reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 36599–36612. Curran Associates, Inc., 2022. doi: 10.52202/068431-2652. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/ed3cd2520148b577039adfade82a5566-Abstract.html.

Shishir G Patil, Huanzhi Mao, Fanjia Yan, Charlie Cheng-Jie Ji, Vishnu Suresh, Ion Stoica, and Joseph E. Gonzalez. The berkeley function calling leaderboard (BFCL): From tool use to agentic evaluation of large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 48371–48392. PMLR, 2025. URL https://proceedings.mlr.press/v267/patil25a.html.

Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1511.05952.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

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, 2024. URL https://arxiv.org/abs/2402.03300.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. ALFWorld: Aligning text and embodied environments for interactive learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=0IOX0YcCdTn.

Richard S. Sutton, David A. McAllester, Satinder P. Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, volume 12, 1999. URL https://proceedings.neurips.cc/paper/1999/hash/464d828b85b0bed98e80ade0a5c43b0f-Abstract.html.

Harsh Trivedi, Tushar Khot, Mareike Hartmann, Ruskin Manku, Vinty Dong, Edward Li, Shashank Gupta, Ashish Sabharwal, and Niranjan Balasubramanian. AppWorld: A controllable world of apps and people for benchmarking interactive coding agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16022–16076, 2024. doi: 10.18653/v1/2024.acl-long.850. URL https://aclanthology.org/2024.acl-long.850/.

Yinjie Wang, Tianbao Xie, Ke Shen, Mengdi Wang, and Ling Yang. RLAnything: Forge environment, policy, and reward model in completely dynamic RL system. In *International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=JLQV6NxS6W.

Shidong Yang, Ziyu Ma, Tongwen Huang, Yiming Hu, Yong Wang, and Xiangxiang Chu. CoEvolve: Training LLM agents via agent-data mutual evolution. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23015–23036. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1055. URL https://aclanthology.org/2026.acl-long.1055/.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. WebShop: Towards scalable real-world web interaction with grounded language agents. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://papers.nips.cc/paper_files/paper/2022/hash/82ad13ec01f9fe44c01cb91814fd7b8c-Abstract-Conference.html.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.03629.

Gyeongtae Yoo, Sanghyeok Park, Soohyuk Jang, Ik-hwan Kim, and Sungroh Yoon. Rollout-level advantage-prioritized experience replay for GRPO, 2026. URL https://arxiv.org/abs/2606.04560.

Yifei Zhou, Andrea Zanette, Jiayi Pan, Sergey Levine, and Aviral Kumar. ArCHer: Training language model agents via hierarchical multi-turn RL. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62178–62209, 2024. URL https://proceedings.mlr.press/v235/zhou24t.html.

Appendix Contents

Supplementary material			Page
A	**Technical Details**		13
	A.1	Execution contract and feature encoding	13
	A.2	Linear estimator and efficient evaluation	13
	A.3	Budget allocation and training procedure	14
B	**Experimental Details**		16
	B.1	Preliminary-study setup and metrics	16
	B.2	Benchmarks, splits, and native metrics	16
	B.3	Environment changes and the restoration contract	17
	B.4	Replay baselines and system comparisons	17
	B.5	Budgets, checkpoint selection, and statistical reporting	18
	B.6	Backbones and implementation defaults	19
	B.7	Mechanism diagnostics and ablations	20
	B.8	Training-time gradient phase protocol	21
	B.9	Native evaluation and resource conventions	21
	B.10	Within-label misspecification protocol	22
C	**Additional Experimental Results**		23
	C.1	Preliminary-study diagnostics	23
	C.2	Controlled adaptation and terminal performance	23
	C.3	Native-protocol comparisons	23
	C.4	Transfer to held-out tool-use tasks	24
	C.5	Component, mechanism, and stability analyses	24
	C.6	Gradient correction over training	27
	C.7	Numerical sensitivity to within-label changes	27
D	**Proofs**		28
	D.1	Population assumptions	28
	D.2	Population least-change characterization	28
	D.3	Identification from paired probes and exact feature residuals	29
	D.4	Trajectory change of measure and gradient consistency	33
	D.5	Policy-gradient identity and correction error	39
	D.6	Quantitative consequences of uncorrected reliability drift	40
	D.7	Necessity, opposite gradients, and equality cases	44
	D.8	Information lost by the fixed features	45
	D.9	Missing support and the exact reweighting limit	46

A TECHNICAL DETAILS

A.1 EXECUTION CONTRACT AND FEATURE ENCODING

A.1.1 RESTORATION, OUTCOMES, AND REPLAY RECORDS

The environment exposes a common serialization of the complete state and immutable version identifiers. A snapshot contains the application database, interpreter variables, session state, and any clock value that affects the next transition. Restoration reconstructs this state in an isolated process. Future random draws are generated independently after restoration. Both source and current copies must be restored successfully before either probe action is executed; restoration rejection depends only on the pre-execution record. Inaccessible anchors are excluded from the sampling pool. Restoration checks and reconstruction calls are charged to the interaction ledger.

An affected action invokes one tool through a reliability wrapper. The wrapper emits $y = 0$ when it returns the designated failure response and $y = 1$ when it executes the underlying tool. Application-level errors returned during normal execution remain in class $y = 1$. The wrapper records this execution label directly, separately from the reward and response text. Equation (2) assumes that the wrapper's class probabilities depend only on tool identity, and that the complete within-class state/observation law is unchanged across versions. This specification covers one affected call per action, with reliability determined by tool identity. Termination and actions without an affected call use their common, version-independent kernel.

Each replay record stores its task identifier, version, full action tokens, behavior log probabilities, observations, outcome labels, terminal reward, and pre-execution snapshots. A valid changed transition has a recognized tool identifier, an explicit binary label, and a usable snapshot. A record with missing action tokens, missing reward, nonfinite behavior log probabilities, or an invalid changed transition contributes zero to the replay sum and remains in its denominator. Its score need not be evaluated. Unexpected execution failures outside the wrapper contract invalidate the entire probe pair; both branches' incurred costs remain charged.

Behavior and current policies use fixed positive-temperature softmax sampling with the same parameter-independent token masks, tokenizer, and generation limit. Action likelihoods include reasoning, tool arguments, and termination tokens. Environment versions are held fixed for the complete fresh batch, all probes, and the subsequent policy step. A version change starts a new round and clears the probe cache, fits, and allocation cursor.

A.1.2 FEATURE ENCODING AND TASK FOLDS

The shared reliability-wrapper registry determines the affected-tool manifest. Sort its canonical `application.operation` identifiers lexicographically by UTF-8 bytes and assign indices $k = 1, \ldots, K$. Coordinate $2k - 1$ denotes wrapper failure and coordinate $2k$ normal execution. The feature vector has exactly one unit entry and dimension $2K$; every other coordinate is zero. The encoding uses only tool identity and the binary outcome label. A tool outside this manifest has ratio one only if the version specification declares its kernel unchanged. An unrecognized changed tool makes the record invalid.

The task fold is the least significant bit of the SHA-256 digest of its canonical UTF-8 task identifier. This assignment is fixed before training. Fit (v, f) uses only probes from tasks in fold $1 - f$ and scores only tasks in fold f. All trajectories from a task share its fold. Counts $N^{-}_{v,f,k,0}$ and $N^{-}_{v,f,k,1}$ are computed from valid pairs in that fit, never borrowed from the other fold. Thus a tool absent from the opposite fold is ineligible even when it appears elsewhere in the buffer. Under Equation (2), tool identity is the relevant anchor class; no finite partition of the full state space is required.

A.2 LINEAR ESTIMATOR AND EFFICIENT EVALUATION

For a fixed fit, let $r_w = 1 + (\phi - \widehat{\mu})^\top w$. Expanding the empirical least-squares risk gives

$$\frac{1}{2m}\sum_j r_w(u_j, z_j^-)^2 - \frac{1}{m}\sum_j r_w(u_j, z_j^+) + \frac{\lambda}{2}\|w\|_2^2 = \frac{1}{2}w^\top \widehat{C} w - \widehat{\Delta}^\top w + \frac{\lambda}{2}\|w\|_2^2 - \frac{1}{2}. \quad (11)$$

Its Hessian $\widehat{C} + \lambda I$ is positive definite for $\lambda > 0$, so Equation (6) has a unique solution. Centering fixes the empirical source mean of the raw ratio at one, while the ridge shrinks the correction in weakly supported directions.

Let $\widehat{\mu}^{+} = \widehat{\mu} + \widehat{\Delta}$ be the target frequency vector. The one-hot covariance gives an exact evaluation formula for every coordinate j:

$$\widehat{r}_j = \frac{\widehat{\mu}_j^{+} + \lambda\gamma}{\widehat{\mu}_j + \lambda}, \qquad \gamma = \frac{\sum_{j=1}^{2K} \widehat{\mu}_j^{+}/(\widehat{\mu}_j + \lambda)}{\sum_{j=1}^{2K} \widehat{\mu}_j/(\widehat{\mu}_j + \lambda)} > 0. \tag{12}$$

To derive it, set $t = \widehat{\mu}^{\top}\widehat{w}$. Coordinate j of the normal equation is $(\widehat{\mu}_j + \lambda)\widehat{w}_j - \widehat{\mu}_j t = \widehat{\mu}_j^{+} - \widehat{\mu}_j$. Substitution into $\widehat{r}_j = 1 + \widehat{w}_j - t$, followed by $\sum_j \widehat{\mu}_j \widehat{r}_j = 1$, yields Equation (12) with $\gamma = 1 - t$. Both frequency vectors are nonnegative and sum to one, proving strict positivity and normalization. Evidence exclusions and trajectory capping are applied to these positive, normalized ratios.

Equation (12) is the numerical implementation of the covariance solve. Accumulating counts costs $O(m + K)$ time and $O(K)$ memory; evaluating all cell ratios costs $O(K)$, and scoring a transition is a constant-time lookup. With V active old versions, there are at most $2V$ fits and $O(VK)$ moment storage. Snapshot storage is separate. Empty counts and nonfinite evaluations make the fit unavailable. For an unobserved tool, both source and target coordinates vanish and the raw ratio equals γ; the evidence rule sets its usable replay factor to zero.

The population least-change characterization and its proof are in Appendix D.2; practical correction error is analyzed in Appendix D.5.

A.3 Budget allocation and training procedure

A.3.1 Cost reservation and cyclic allocation

A charged call is one environment execution, including executions used by initialization and restoration. Non-call restoration work is recorded separately as wall time. The execution interface enforces integer caps $L_F \geq 1$ per complete fresh trajectory and $L_P \geq 1$ per probe pair; both include failed calls and the relevant initialization/restoration calls. For example, when each branch needs at most L_R restoration calls and one tool execution, $L_P = 2(L_R + 1)$. The horizon and fixed timeout/termination rules bound L_F. The ledger N_{call} increases by actual incurred calls, including invalid attempts.

Compute Equation (7) before collecting the fresh batch. Its reservation satisfies $N_F + m_n L_P \leq N_{\max} - N_{\mathrm{call}}$ and $m_n L_P/(N_F + m_n L_P) \leq \alpha$. The latter constrains the reserved costs; shorter episodes can change the realized fraction. Unused reservations return to the remaining budget for the next round; the current round's quota stays fixed. If fewer than N_F calls remain, training stops before collecting another batch. At $\alpha = 0$, no new probes are collected.

For an old version v, scored fold f, and affected tool k, form a pool of usable pre-execution anchors in the current buffer with version v, tool k, and task fold $1 - f$. The triple (v, f, k) is eligible exactly when this pool is nonempty, version v is executable, and the buffer contains a task in fold f using tool k in version v. Eligibility depends solely on pre-execution metadata. Freeze the eligible set and its pools at the start of the round; fresh trajectories are appended only after the update.

Sort eligible triples by version-identifier bytes, then fold, then tool index. The persistent cursor stores the last attempted triple. Start at its first strict lexicographic successor in the current set, wrapping to the smallest triple when needed; an empty cursor starts at the smallest triple. Each attempt samples one anchor uniformly with replacement from its pool and advances the cursor once, regardless of success. For J eligible triples, a quota m_n assigns $\lfloor m_n/J \rfloor$ attempts to each, with one additional attempt for the first $m_n \bmod J$ triples in cyclic order. The empty-set case executes zero attempts. Uniform draws and independent branch randomness use the run's fixed random seed.

Both restoration branches must succeed before executing a pair. A valid pair contributes one source and one target outcome to the same fit (v, f). All other attempts contribute no moments. Valid counts accumulate while the target version is fixed, including when their anchor records later leave the FIFO buffer. Pooling these counts is valid because class probabilities depend only on tool identity. Refit only systems whose counts changed. The controls are $\alpha \in [0, 1)$, $\lambda > 0$, and $c \geq 1$; the batch size, FIFO capacity, step sizes, execution caps, and seed specify the shared training protocol.

A.3.2 COMPLETE ALGORITHM

The buffer capacity is an integer $M_{\mathrm{buf}} \geq B$. Uniform old-trajectory sampling always uses replacement. Rejected records count toward the sampled batch size. A nonempty buffer keeps the same fresh coefficient even when every old record is rejected.

Algorithm 1 ESR: environment-corrected trajectory replay

Require: Policy θ; empty FIFO buffer $\mathcal{B}$ of capacity M_{buf}; task/tool manifests; batch size B, budget $N_{\max}$, execution caps L_F, L_P, seed; probe fraction α, ridge λ, cap c, step sizes ϵ_n.

$b \leftarrow 0$; $N_{\mathrm{call}} \leftarrow 0$; clear probe counts, fits, and cursor.
while $N_{\max} - N_{\mathrm{call}} \geq BL_F$ **do**
  Freeze $\theta_n \leftarrow \theta$ and target version $v_\star$ for this round.
  On a version change, clear all probe counts, fits, and the cursor.
  Freeze eligible triples and anchor pools; compute m_n by Equation (7).
  Collect B fresh trajectories $\mathcal{F}_n$; charge actual calls to N_{call}.
  if the eligible set is nonempty **then**
    for m_n cyclic attempts **do**
      Sample an opposite-fold anchor in the next triple and advance the cursor.
      Restore both copies, then execute with independent randomness; charge all calls.
      Add both outcomes to the fit's counts exactly when the pair is valid.
    end for
  end if
  Recompute changed fits using Equation (12).
  if $\mathcal{B} = \varnothing$ **then**
    $\mathcal{L}_n(\theta) \leftarrow -B^{-1} \sum_{i \in \mathcal{F}_n} \mathrm{sg}(R_i - b) S_i(\theta)$.
  else
    Sample B old trajectories $\mathcal{H}_n$ uniformly with replacement.
    Apply record validity and both-source-outcome checks to every changed transition.
    Compute trajectory weights by Equation (8) and loss by Equation (9).
  end if
  $\theta \leftarrow \theta_n - \epsilon_n \nabla \mathcal{L}_n(\theta_n)$; $b \leftarrow B^{-1} \sum_{i \in \mathcal{F}_n} R_i$.
  Append $\mathcal{F}_n$ to $\mathcal{B}$; evict the oldest excess records.
end while
return θ, buffer, and call/evidence/weight ledger.

Positive transition factors are accumulated in log space; a zero factor sets the trajectory weight to zero immediately. There is no positive floor. A numerical failure makes its fit unavailable and zeros the affected replay contributions. Neither weights nor rewards receive gradients, and the fit is frozen for the policy step. The ledger records source/current probe calls, fresh calls, restoration calls, invalid-pair counts, eligible tool cells, and zero/capped trajectory weights. For each old batch, report $\mathrm{ESS} = (\sum_{i \in \mathcal{H}_n} \widetilde{W}_i)^2 / \sum_{i \in \mathcal{H}_n} \widetilde{W}_i^2$, with value zero when the denominator vanishes. ESS summarizes weight concentration; the loss remains as defined in Equation (9).

B EXPERIMENTAL DETAILS

We evaluate the least-squares ESR estimator in Section 4 through controlled adaptation, native-protocol comparisons, and held-out transfer. This appendix specifies the benchmarks, training configurations, and resource accounting; Appendix C.5 examines component effects, gradient correction, stability, and probe cost.

B.1 PRELIMINARY-STUDY SETUP AND METRICS

B.1.1 CONTROLLED RELIABILITY DRIFT

The source environment v^- and current environment v^+ share the same state and action spaces, reward function, tool semantics, and conditional response distribution given a tool outcome. Only the probabilities of normal tool execution change. For $K = 8$ affected tools, let p_k^- and p_k^+ denote these probabilities and define

$$\delta = \frac{1}{K} \sum_{k=1}^{K} |p_k^+ - p_k^-|. \tag{13}$$

The intervention balances improvements and degradations across tools. We cross seven drift levels, $\delta \in \{0, 0.03, 0.06, 0.10, 0.14, 0.20, 0.28\}$, with eight affected-call counts, $n_{\text{tool}} \in \{1, 2, 3, 4, 6, 8, 10, 12\}$, and use 24 random task/environment seeds per setting.

The policy π_θ remains frozen. Each seed uses $N_- = 256$ source trajectories and an independent batch of $N_+ = 2048$ fresh current-environment trajectories. The latter estimates a reference gradient for this diagnostic. The behavior policy satisfies $\beta = \pi_\theta$, so every policy likelihood ratio equals one; replay gradients receive no environment correction.

B.1.2 GRADIENT METRICS AND AGGREGATION

For environment $e \in \{-, +\}$, we estimate

$$\widehat{\mathbf{g}}_e = \frac{1}{N_e} \sum_{i=1}^{N_e} (R_i - b) \sum_{t<T_i} \nabla_\theta \log \pi_\theta(a_{i,t} \mid h_{i,t}), \tag{14}$$

where R_i is trajectory return, T_i is trajectory length, $h_{i,t}$ is the history before action $a_{i,t}$, and b is a fixed scalar baseline shared by both environments. We compute cosine alignment, relative error, and a reversal indicator separately for each seed:

$$A_{\text{grad}} = \frac{\widehat{\mathbf{g}}_-^\top \widehat{\mathbf{g}}_+}{\|\widehat{\mathbf{g}}_-\|_2 \|\widehat{\mathbf{g}}_+\|_2}, \tag{15}$$

$$E_{\text{grad}} = \frac{\|\widehat{\mathbf{g}}_- - \widehat{\mathbf{g}}_+\|_2}{\|\widehat{\mathbf{g}}_+\|_2}, \qquad I_{\text{rev}} = \mathbf{1}[A_{\text{grad}} < 0]. \tag{16}$$

Reported means average these seed-level metrics; the reversal fraction is $\sum_s I_{\text{rev},s}/24$. A reversal denotes opposition to the estimated fresh gradient. Across the $56 \times 24 = 1344$ seed-level observations, the norm ratio $q = \|\widehat{\mathbf{g}}_-\|_2 / \|\widehat{\mathbf{g}}_+\|_2$ and signed angle φ satisfy $\cos\varphi = A_{\text{grad}}$ and $E_{\text{grad}}^2 = q^2 + 1 - 2qA_{\text{grad}}$ for each observation.

B.2 BENCHMARKS, SPLITS, AND NATIVE METRICS

Primary adaptation benchmark. AppWorld provides executable applications, persistent state, and programmatic task verification (Trivedi et al., 2024). We use the official Train/Dev/Test-N/Test-C partitions (105/60/168/417 tasks), and report Test-N and Test-C separately. TGC is the fraction of tasks passing all tests; SGC requires success on every task variant within a scenario. We compute SGC within each evaluation repetition and then average repetitions. Training trajectories, probe anchors, fitting data, and task synthesis use Train exclusively; Dev selects configurations and checkpoints. Both test suites are reserved for evaluation. For fold hashing, the canonical task identifier is its scenario/template key; individual episode IDs remain distinct. All variants retain the same fold and official split assignment. Source and target versions share initial task states and evaluation tests.

Additional environments. ALFWorld and WebShop provide direct overlap with GiGPO and DreamGym (Feng et al., 2025; Chen et al., 2026). We use ALFWorld's 3,553 training tasks and 140 seen-development tasks, with 134 unseen tasks as the primary held-out evaluation. WebShop uses its released training, development, and test manifests, frozen with the product catalog and simulator. We report the official continuous score and binary success separately. Controlled training uses terminal success in $\{0, 1\}$ on every benchmark; native reward shaping is retained only in explicitly labeled reference runs. Results are reported separately for each benchmark.

Held-out tool-use transfer. BFCL evaluation uses V3 Multi-Turn Base, following the AppWorld-to-BFCL comparison in CoEvolve (Yang et al., 2026; Patil et al., 2025). We use its released state/response checkers and report accuracy for that subset. AppWorld Dev selects the trained checkpoint, and BFCL is used exclusively for evaluation.

B.3 Environment changes and the restoration contract

Controlled overlapping-support changes. We implement version changes through the reliability mixture in Eq. (2). Before an affected execution, an independent Bernoulli variable selects normal execution or a marked infrastructure failure. Failure leaves the environment unchanged and emits a fixed diagnostic; normal execution uses the unchanged benchmark engine. Ordinary API exceptions and unsuccessful task actions remain normal-branch outcomes. Initial states, prompts, tool schemas, rewards, stopping rules, and the conditional response law within each branch are held fixed. Version identifiers and probe state are kept as metadata outside the policy prompt.

Each generated turn permits at most one affected execution. In AppWorld, the affected unit is the top-level code-execution service; a code block may contain multiple application API calls within its unchanged normal branch. Its default manifest therefore has $K = 1$. ALFWorld/WebShop use their fixed top-level action classes. The $K = 8$ synthetic diagnostic remains a separate setting. Multiple affected services per turn require a corresponding extension of the trajectory factorization.

The main controlled comparison changes the normal-execution probability from 0.7 to $0.7 + \sigma_k 0.2$ at 40% of the 10^6-call budget. AppWorld uses $\sigma_1 = -1$; ALFWorld/WebShop alternate signs in their fixed action-class manifests. Both branches retain nonzero support. The switch occurs only between complete training rounds, and its realized call count is recorded. Every method receives the same version-change signal; only the exact-ratio diagnostic uses the generating probabilities. Supplementary static controls on Qwen2.5-7B/AppWorld keep reliability at 0.7: one has no version signal, and one changes only the version identifier at 40%. The latter clears ESR's fits and incurs ordinary probe costs despite an unchanged transition law.

Restoration and evidence checks. An anchor includes the database, interpreter/session state, persistent files, clocks, and all other variables affecting future transitions. Source and target probe executions start from independently restored copies of that anchor and use independent future randomness. An environment enters the ESR adaptation panel only after complete-state restoration passes state hashing, deterministic replay where applicable, and branch-conditioned response checks. Failed restoration attempts still incur their actual costs.

Probe fits use the task-level two-fold partition and opposite-fold anchors specified in Appendix A.1. Only correctly labeled pairs update the moments. Both source outcomes must be observed for every changed tool used in a scored trajectory; an unseen tool or a failed evidence check gives that trajectory zero replay weight. Version switches clear all fits and counts. Unchanged/current-version transitions have ratio one.

Table 6 uses separate 10^6-call ledgers for ALFWorld and WebShop, the fixed tool-class sign manifest with $\delta = 0.2$, and common rewards, decoding, tasks, and initialization. All entries are controlled-protocol measurements.

B.4 Replay baselines and system comparisons

Common-trainer replay comparisons. All variants below use the same initialization, rollout representation, binary reward, FIFO capacity, and Eq. (9) where applicable. Fresh-only uses its fresh mean with coefficient one. Buffer-flush removes previous-version records on every switch. Naive uniform replay (Arnal et al., 2026) sets all valid replay weights to one; policy-only replay retains the

Table 4: System comparison protocols. Native-method reruns use the stated common backbone; published results retain their original data, interfaces, and budgets.

Method	Task and model	Additional resources
GiGPO (Feng et al., 2025)	WebShop; Qwen2.5-7B	Grouped rollouts and step-level credit assignment
DreamGym (Chen et al., 2026)	WebShop; Qwen2.5-7B and Llama-3.1-8B	Experience-model training, demonstrations, synthetic rollouts, real adaptation
RLAnything (Wang et al., 2026)	ALFWorld; Qwen2.5-7B	Learned reward model and task-generation model
CoEvolve (Yang et al., 2026)	AppWorld; Qwen2.5-7B and Qwen3-4B	Teacher calls, task synthesis, and execution-based validation

complete-trajectory behavior-policy ratio but sets the environment factor to one. An age-only ablation multiplies the policy-only weight by $\exp(-a_i/\tau)$, where a_i counts update steps since collection.

PER-style and FreshPER sampling (Schaul et al., 2016; Ma et al., 2026) use $P(i) \propto p_i^{0.6}$ with reward priority $p_i = |R_i| + 10^{-6}$; FreshPER additionally multiplies p_i by $\exp(-a_i/\tau)$. We fix $\tau = 500$. Following the reported FreshPER variants, we evaluate sampling-bias correction both disabled and enabled with exponent 0.4 and buffer-wide maximum normalization. This correction is distinct from policy importance weighting. Both variants retain policy correction and use a unit environment factor in the common trainer. Buffer size and replay volume are matched here.

Dynamics-aware controls. A DARC-style adaptation (Eysenbach et al., 2021) adds $\eta \sum_t \log \widehat{\rho}^E_{i,t}$ to the replay return, while retaining policy correction. It uses the same paired evidence and estimator as ESR, isolating additive reward correction from multiplicative trajectory correction. Ineligible records contribute zero before any logarithm is evaluated. We tune $\eta \in \{0, 0.01, 0.1, 1\}$. A direct count-ratio baseline estimates $p_v(y \mid k)$ and $p_\star(y \mid k)$ from the same probes, with symmetric pseudocount $1/2$, and uses their ratio in ESR's trajectory weight. Its source-evidence gate still requires observations of both source outcomes. In the tool/label mixture, this baseline provides a direct comparison with ridge fitting. Exact known ratios provide a diagnostic reference whose measured performance also depends on capping and finite-buffer training.

Task-level and recent system comparisons. Table 4 summarizes the task and model settings for system comparisons. The native optimization baselines are PPO, RLOO, and GRPO (Schulman et al., 2017; Ahmadian et al., 2024; Shao et al., 2024); a frozen ReAct agent (Yao et al., 2023) provides the pre-training reference. The native result tables reproduce published rows with their original optimization objectives, optimizer, critic, grouping, data, and resource settings.

PPO/RLOO/GRPO are shared agent baselines in GiGPO, and GRPO also appears in CoEvolve. DreamGym's real-transition count excludes offline demonstrations and auxiliary-model work, which are disclosed with its native results. RLAnything changes rewards and tasks, while CoEvolve changes the training-task distribution. These system comparisons reflect their respective task distributions q, rewards, and training resources. External data access and evaluation protocols are reported alongside the native results.

PER/FreshPER use the common-trainer adaptations in Appendix B.4. Both retain policy importance weighting; sampling-bias correction is disabled except in "+ sampling IS" (exponent 0.4, buffer-wide maximum normalization). Count-ratio correction shares ESR's probes, evidence gate, cap, and replay volume. DARC-style uses the same estimator and adds log ratios to the return, allowing comparison of additive and multiplicative correction.

B.5 Budgets, checkpoint selection, and statistical reporting

For controlled adaptation, the training budget is 10^6 charged training environment operations per benchmark, method, and seed, including source-version data collection. We count rollout executions, both sides of every paired probe, execution-based initialization/restoration, and failed calls in a common ledger; non-call restoration work is timed separately. We also report nested application API operations. Each benchmark fixes the atomic-operation definition and conservative L_F, L_P bounds before training. ESR uses Eq. (7); α bounds the reserved probe fraction. The default 10% bound

Table 5: Common-trainer configuration and hardware. Hyperparameters are selected using development data only.

Item	Configuration
Trainable parameters	Full model; bfloat16 computation, float32 reductions; sharded training
Update	One SGD step, no momentum/weight decay; default $\epsilon_n = 10^{-5}$
Fresh/replay batch	$B = 32$ fresh plus 32 replay trajectories; replay with replacement
Buffer	FIFO, 4,096 complete trajectories; no success-only filtering
ESR parameters	$\alpha = 0.1$, $\lambda = 10^{-3}$, $c = 10$; $2K$ unscaled one-hot features
Training decoding	Temperature 1, top-$p = 1$, top-k disabled; fixed legal-token masks
AppWorld limits	50 generated turns; 16,384-token input and 2,048 new tokens per turn
ALFWorld/WebShop limits	50/15 turns; 2,048/4,096 input tokens and 512 new tokens per turn
Hardware	8×H100 80GB; 64 CPU cores; 512GB host memory

applies to reserved costs; realized probe usage depends on actual execution costs. Methods without probes may spend their saved budget on fresh interaction. Optimizer-step counts may differ under the same environment budget.

The supplementary resource comparison uses Qwen2.5-7B/AppWorld on identical hardware and records fresh, probe, and restoration calls, measured GPU-hours, and wall time. GPU-hours include rollout inference, likelihood recomputation, fitting, and model updates. Each execution is charged once; failed executions remain in the category that incurred them. Evaluation uses a separate budget, identical across methods, and its data are reserved for scoring. Checkpoints are evaluated every 5% of the training-call budget and immediately before/after the switch. The pre-update policy evaluated in the new environment defines zero-adaptation performance.

The primary statistic is trapezoidal target-version success AUC, interpolated and normalized within the post-switch call interval. Terminal performance uses the last budget-feasible checkpoint. Development data select configurations, and the terminal checkpoint is fixed by the call budget. In the static and null-switch controls, terminal TGC is compared with policy-only replay using a one-percentage-point non-inferiority margin and a one-sided 95% lower confidence bound, separately for Test-N and Test-C.

Controlled comparisons use seeds 0, 1, and 2, sharing task/schedule seeds across methods, and report mean and sample standard deviation. The supplementary Qwen2.5-7B/AppWorld comparison reports per-seed AUCs and paired 95% bootstrap intervals for ESR minus fresh-only, policy-only, and count-ratio correction. Bootstrap resampling is paired across methods and clustered by seed and scenario/template, keeping repeated trials within their task cluster. The intervals summarize uncertainty across three training seeds, providing limited resolution for small differences. Configuration selection uses development performance; exact wrapper ratios and test-gradient diagnostics are reserved for analysis.

Table 1 uses $\delta = 0.2$ degradation and one switch after 40% of the 10^6-call budget. AUC integrates from the pre-update policy evaluated in the new environment to the last budget-feasible checkpoint. Test-N/Test-C use separate TGC curves. The ledger includes collection, paired probes, and charged restoration (Appendix B.5).

B.6 Backbones and implementation defaults

The exact checkpoint names are `Qwen/Qwen2.5-7B-Instruct`, `Qwen/Qwen3-4B`, and `meta-llama/Llama-3.1-8B-Instruct`. Qwen2.5-7B is the common backbone across GiGPO, DreamGym, CoEvolve, and RLAnything; Qwen3-4B follows the second CoEvolve backbone, while Llama-3.1-8B tests another model family. All three run the AppWorld replay comparison. The additional environment and system comparisons use the model–benchmark combinations in Table 4.

SGD implements the update in Section 4.3. Native algorithm runs retain their optimizer, critic, grouping, and clipping, starting from the published settings and receiving the same tuning allowance.

Training uses complete-trajectory scores without length normalization, self-normalized weights, or group-standardized advantages. Invalid or unsupported replay records keep their place in the denominator, and a nonempty buffer retains fresh coefficient $1/2$, even when every old weight is zero.

Native starting configurations include GiGPO's actor learning rate 10^{-6} and group size eight; CoEvolve's group size eight, learning rate 10^{-6}, and KL coefficient 10^{-3}; and RLAnything's ALFWorld limits of 40 training/60 evaluation turns (Feng et al., 2025; Yang et al., 2026; Wang et al., 2026). Their different limits and rewards require separate native-protocol rows. Matched-budget reruns use common limits and an identical evaluator for every method; deviations from native settings are reported with the results.

Sampling probabilities cover every generated token—reasoning text, arguments, and sampled stop tokens—under the collection decoder. The training decoder avoids policy-dependent support truncation. We fix Qwen3 to non-thinking mode for the common comparison. Prompt templates and deterministic history truncation are identical across methods for a given model, and stored model inputs allow exact likelihood recomputation. Hitting turn, token, or execution limits is scored by the frozen failure rule. The ALFWorld/WebShop limits follow the GiGPO evaluation protocol; published results with different limits or sampling are reported as native-protocol references. Controlled evaluation uses greedy action decoding with four independent environment trials per task; native reference evaluations retain their reported decoder. Both use identical decoding across methods within a table.

The default ESR configuration is $\alpha = 0.1$, $\lambda = 10^{-3}$, $c = 10$, and $\epsilon_n = 10^{-5}$. These values remain fixed across held-out tasks and the three backbones. Configuration selection uses development data only; supplementary component variants each change one stated factor. Reproducibility records identify software/container revisions, task and sign manifests, checkpoint hashes, and the selected configuration.

B.7 Mechanism diagnostics and ablations

Component effects and probe cost. Supplementary training comparisons use Qwen2.5-7B/AppWorld with the main $0.7 \to 0.5$ switch, 10^6 charged calls, and three seeds. Policy-only, FreshPER, count-ratio, and ESR provide the reference results. The additional variants are age-only weighting ($\tau = 500$), lower probe reservation ($\alpha = 0.05$), weaker ridge ($\lambda = 10^{-4}$), removal of the trajectory cap, and removal of the two-outcome source-evidence gate. Other factors, including the 4,096-trajectory buffer and SGD update, remain fixed. The last variant extrapolates to missing source outcomes while still excluding invalid records and wholly unprobed tools.

We summarize post-switch replay coverage, cap frequency, and normalized effective sample size,

$$\mathrm{nESS} = \frac{(\sum_i \widetilde{W}_i)^2}{B \sum_i \widetilde{W}_i^2}, \tag{17}$$

with nESS zero for an all-zero batch. Coverage and cap frequency use all B sampled replay records as their denominator; cap frequency counts strictly capped, eligible weights. Each statistic is averaged over post-switch update batches within a seed and then across seeds. Removing the cap sets the cap frequency to zero by definition. Probe reservation and actual probe usage are reported separately in Table 12.

Frozen-policy correction diagnostic. We freeze Qwen2.5-7B at each seed's pre-switch checkpoint and collect diagnostic source trajectories with that policy, so the policy ratio equals one. Paired fits use opposite-fold training anchors; held-out diagnostic tasks are reserved for evaluation. We compare policy-only, count-ratio, ESR, and exact uncapped environment ratios on the same source records under $0.7 \to 0.5$. Local log-ratio error is the mean absolute error on the common ESR-eligible records before trajectory capping. Gradient relative error and cosine compare the replay gradient, including ESR's gate and cap, with an independently sampled current-environment reference. Zero-weight records remain in the gradient denominator. Every method and reference uses the same fixed reward baseline and complete-trajectory score; the diagnostic evaluates the replay gradient before multiplication by the fresh/replay mixing coefficient.

Table 10 separates trajectories with 1–4 and at least 5 affected executions. Source and target gradients use the same stratum definition; gradients are probability-weighted contributions to the full task objective, using the total sample count across strata as the denominator. Three independent target-

reference batches per seed quantify reference Monte Carlo variability; their pairwise comparisons form the fresh-reference row. We average reference-batch metrics within each seed before computing the three-seed mean and sample standard deviation. Diagnostic sample counts are recorded with the table. Relative error and cosine are marked undefined when the reference norm is zero. Diagnostic executions are accounted for separately and used exclusively for analysis.

B.8 Training-time gradient phase protocol

Section 5.4 uses Qwen2.5-7B-Instruct on AppWorld with the $0.7 \to 0.5$ switch at 40% of the 10^6-call training budget. For each of three training seeds, the policy is frozen for diagnostic scoring at 11 evenly spaced post-switch checkpoints, including the first after the switch and the final budget-feasible checkpoint. The policy and FIFO buffer continue to evolve between checkpoints. At each checkpoint, 32 matched minibatches of 64 trajectories are sampled from the current buffer. The same records receive (i) policy likelihood-ratio correction only, (ii) full ESR, including fitted environment factors, the evidence rule, and trajectory cap, or (iii) policy likelihood ratios combined with the generating environment ratios, without capping (*Exact ratio*).

An independent batch of 8,192 current-environment trajectories per seed and checkpoint estimates $\mathbf{g}_{\text{fresh}}$. Replay and fresh gradients use the same fixed reward baseline and complete-trajectory score; replay gradients are evaluated before fresh/replay mixing, with zero-weight records retained in the minibatch denominator. These executions are used exclusively for diagnostics and accounted for separately from the training budget. Exact ratios provide a comparison with the fitted correction against this finite-sample fresh reference.

For replay gradient $\mathbf{g}_m$ and the independently estimated current-environment gradient $\mathbf{g}_{\text{fresh}}$, define

$$d_m = \frac{\mathbf{g}_m^\top \mathbf{g}_{\text{fresh}}}{\|\mathbf{g}_m\|_2 \|\mathbf{g}_{\text{fresh}}\|_2}, \qquad s_m = \log_2 \frac{\|\mathbf{g}_m\|_2}{\|\mathbf{g}_{\text{fresh}}\|_2}. \tag{18}$$

The reference lies at $(1, 0)$; $d_m < 0$ indicates directional reversal, and $s_m = \pm 1$ indicates twice or half its norm.

Each of the $3 \times 11 \times 32 = 1056$ matched minibatches contributes one (d_m, s_m) pair per method. Checkpoint means pool all 96 minibatches at that checkpoint; improvement fractions are descriptive summaries over all matched pairs, with the training seed as the independent unit. Figure 5 displays every recorded point and arrows for batch indices $0, 8, 16, 24$ at every seed/checkpoint (132 pairs). Gaussian kernel densities use bandwidth factor 0.22; contours are placed at the 10th, 35th, 60th, and 80th percentiles of density evaluated at observed points. Density evaluation is restricted to the valid cosine domain $[-1, 1]$.

B.9 Native evaluation and resource conventions

AppWorld. Table 2 takes base-model/CoEvolve values from CoEvolve's Table 1 and GRPO Test-N TGC from Table 2 (Yang et al., 2026); unreported metrics are "–". ESR uses the native interface, splits, decoder, and 30-turn/4,096-token limits, distinct from the common trainer's 50/2,048 limits. Checkpoint revision and Qwen3 chat-template mode are also part of the source protocol.

WebShop and ALFWorld. Table 3A separates score from success. GiGPO's two rows divide episode advantages by group standard deviation or by one. The native protocol retains reward shaping, temperature 0.4, 15 turns, 4,096 input tokens, and 512 response tokens (Feng et al., 2025). Panel B retains RLAnything's temperature 0.8, 60 evaluation turns, and 140-seen/134-unseen split (Wang et al., 2026). RLAnything's reward/task models count as system resources; its ablations are distinct from our replay controls.

The ALFWorld panel uses results with matching unseen-task manifests. Each native panel retains its source-specific data access, checkpoint selection, resource budget, and evaluator.

DreamGym and resource accounting. DreamGym's WebShop experience model uses 1,600 human and 2,000 oracle/random trajectories with teacher-produced reasoning (Chen et al., 2026). Its zero-online-transition rows exclude those offline data and auxiliary-model costs; S2R adds real-environment fine-tuning.

B.10 WITHIN-LABEL MISSPECIFICATION PROTOCOL

ESR corrects changes in tool–outcome frequencies while assuming invariant within-label response kernels (Eq. (2)). To describe departures from this assumption, consider the family

$$P_{v^\star}^{(\delta,\kappa)}(z \mid u) = \sum_{y\in\{0,1\}} p_{v^\star,\delta}(y \mid k(u))\big[(1-\kappa)G_y(z \mid u) + \kappa\widetilde{G}_y(z \mid u)\big], \tag{19}$$

where δ denotes reliability drift and κ interpolates between the source and altered within-label kernels. The case $\kappa = 0$ lies in ESR's model class; identifying the altered conditional law when $\kappa > 0$ requires information beyond label frequencies.

Figure 6 illustrates score-level sensitivity using a stochastic score construction on a 16×16 grid, with both axes ranging from 0 to 0.30 in steps of 0.02. Using the cosine coordinate d_m in Eq. (18), the plotted gain is $\Delta_{\cos} = d_{\mathrm{ESR}} - d_{\mathrm{PO}}$.

C Additional Experimental Results

C.1 Preliminary-study diagnostics

For twelve affected calls, mean relative error increases from 0.19 at $\delta = 0$ to 1.83 at $\delta = 0.20$ and 2.45 at $\delta = 0.28$. At the intermediate setting $\delta = 0.20$, mean alignment is -0.15 and 17 of 24 seeds reverse.

In Figure 2a, each ray encodes one seed through its angle φ and length q relative to a horizontal unit reference. The circle marks unit norm; blue and orange distinguish nonnegative and negative alignment. This two-dimensional representation preserves alignment and relative magnitude. The compass shows six drift levels; panel (b) includes the complete grid. Each cell aggregates 24 seeds, with disk area proportional to mean relative error. Drift columns are categorical and aligned across all call-count rows.

The dashed frontier in panel (b) uses the first pair of adjacent tested drift levels whose mean alignments cross 0.5, with linear interpolation in δ. The crossings occur at approximately 0.213, 0.173, 0.135, and 0.127 for 6, 8, 10, and 12 affected calls, respectively. For 1–4 calls, mean alignment stays above 0.5 throughout the tested range.

C.2 Controlled adaptation and terminal performance

ESR exceeds fresh-only by 5.66/4.98 terminal-success points on ALFWorld/WebShop. Against count-ratio correction, AUC gains are 0.76/0.68 points and terminal-success gains are 0.87/0.74 points. These differences summarize the three-seed means.

Table 6: **Controlled adaptation beyond AppWorld.** Qwen2.5-7B-Instruct; three-seed means, with s.d. in subscripts. AUC is normalized success-rate AUC ($\times 100$); SR is terminal success rate. ALFWorld uses the 134 unseen tasks with four evaluation trials per task. For SR, s.d. is the sample standard deviation across the three seed-level success rates. WebShop score and SR use the same terminal checkpoint. Bold/underline denote the highest/second-highest mean.

	ALFWorld		WebShop		
Method	AUC	SR	AUC	Score	SR
Fresh-only	$51.82_{\pm 1.22}$	$59.45_{\pm 1.63}$	$60.74_{\pm 1.08}$	$82.11_{\pm 1.34}$	$70.18_{\pm 1.55}$
Buffer-flush	$53.14_{\pm 1.11}$	$60.95_{\pm 1.45}$	$61.66_{\pm 0.99}$	$83.02_{\pm 1.28}$	$71.36_{\pm 1.42}$
Uniform replay	$46.73_{\pm 1.58}$	$54.85_{\pm 1.89}$	$56.21_{\pm 1.44}$	$77.83_{\pm 1.71}$	$65.42_{\pm 1.88}$
Policy-only replay	$47.28_{\pm 1.44}$	$55.35_{\pm 1.73}$	$56.79_{\pm 1.31}$	$78.41_{\pm 1.60}$	$66.01_{\pm 1.76}$
PER	$46.95_{\pm 1.63}$	$55.04_{\pm 1.87}$	$56.48_{\pm 1.51}$	$78.02_{\pm 1.83}$	$65.76_{\pm 1.92}$
FreshPER	$50.61_{\pm 1.29}$	$58.71_{\pm 1.63}$	$59.81_{\pm 1.20}$	$81.19_{\pm 1.49}$	$68.92_{\pm 1.67}$
FreshPER + sampling IS	$50.88_{\pm 1.22}$	$59.08_{\pm 1.55}$	$60.12_{\pm 1.15}$	$81.48_{\pm 1.42}$	$69.25_{\pm 1.59}$
DARC-style	$54.18_{\pm 1.05}$	$62.38_{\pm 1.38}$	$62.89_{\pm 0.97}$	$84.28_{\pm 1.22}$	$72.64_{\pm 1.41}$
Count-ratio correction	$\underline{56.07_{\pm 0.89}}$	$\underline{64.24_{\pm 1.21}}$	$\underline{64.76_{\pm 0.83}}$	$\underline{85.71_{\pm 1.08}}$	$\underline{74.42_{\pm 1.29}}$
ESR (ours)	$\mathbf{56.83_{\pm 0.82}}$	$\mathbf{65.11_{\pm 1.13}}$	$\mathbf{65.44_{\pm 0.78}}$	$\mathbf{86.38_{\pm 1.01}}$	$\mathbf{75.16_{\pm 1.21}}$

C.3 Native-protocol comparisons

AppWorld. Test-N/Test-C TGC gains over CoEvolve are 2.38/1.92 points on Qwen2.5-7B and 2.39/2.15 points on Qwen3-4B; SGC gains are 1.79/0.72 and 1.79/1.44 points, respectively. Evaluation conventions and source attribution are given in Appendix B.9.

WebShop and ALFWorld. ESR improves over GiGPO without standardization by 1.1 score points and 1.5 success points. On ALFWorld, ESR gains 2.3/2.2 success points over RLAnything. ESR reports three-seed means and s.d.

DreamGym and resource accounting. ESR reaches 75.4/75.8 success on Qwen2.5-7B/Llama-3.1-8B, versus 73.7/75.0 for the strongest reported DreamGym variants. ESR's 80K budget counts

Table 7: **WebShop success under the DreamGym protocol.** Published baselines are reproduced from Table 1 of Chen et al. (2026); no standard deviations accompany these point estimates. "Real transitions" counts the source's online real-environment data; offline data and auxiliary-model costs are described below. All rows retain the source evaluation protocol; ESR's budget uses a different counting convention.

Method	Real transitions	Qwen2.5-7B	Llama-3.1-8B
GRPO (Shao et al., 2024)	80K	66.1	65.0
GRPO + DreamGym	0	68.3	63.9
GRPO + DreamGym-S2R	5K	72.1	75.0
PPO (Schulman et al., 2017)	80K	68.1	64.2
PPO + DreamGym	0	65.0	58.1
PPO + DreamGym-S2R	5K	73.7	63.9
ESR (ours)	80K†	**75.4**	**75.8**

† ESR's 80K online execution budget counts both fresh and probe executions; DreamGym reports online real-environment transitions.

fresh and probe executions jointly. The point estimates compare native-task performance under these method-specific resource conventions.

C.4 Transfer to held-out tool-use tasks

Table 8 uses CoEvolve's Tables 2 and 6 (Yang et al., 2026). Its AppWorld-trained Qwen3-4B scores 45.00 on BFCL V3 Multi-Turn Base; the separate 63.00 result uses BFCL-domain training. ESR reaches 47.50 without BFCL training or probes, improving by 2.50 points over CoEvolve and 21.00 over the base model. Its AppWorld value repeats Table 2.

Table 8: **AppWorld-to-BFCL transfer with Qwen3-4B.** Published baselines from Tables 2 and 6 of Yang et al. (2026); evaluation uses its original protocol. The BFCL column reports V3 Multi-Turn Base accuracy.

Method	AppWorld Test-N TGC	BFCL V3 Base
Base model	16.67	26.50
GRPO (Shao et al., 2024)	28.57	–
CoEvolve (AppWorld-trained)	35.71	45.00
ESR (ours; AppWorld-trained)	**38.10**	**47.50**

C.5 Component, mechanism, and stability analyses

C.5.1 Component ablations

Tables 9–12 use Qwen2.5-7B-Instruct/AppWorld under Appendix B. Full ESR has the highest mean AUC among the component variants. Removing the trajectory cap lowers Test-N/Test-C AUC by 0.89/0.61 points and nESS from 0.63 to 0.44. Lower probe reservation reduces coverage, while weaker ridge or removing the two-outcome gate increases cap frequency and lowers nESS. These three-seed comparisons associate useful replay with both evidence coverage and control of weight concentration.

Table 9: **Component ablations.** $0.7 \to 0.5$ at 40% of 10^6 charged calls; AUCs are three-seed means with sample s.d. in subscripts (0–100 scale). N/C: Test-N/Test-C. Replay diagnostics are means over post-switch batches and seeds. Policy-only, FreshPER, count-ratio, and full ESR reuse Table 1. Bold/underline rank AUC means.

	TGC AUC ↑		Replay diagnostics		
Method / changed component	N	C	Cov. (%)	Cap (%)	nESS
Reference methods					
Policy-only replay	$19.16_{\pm 1.12}$	$6.58_{\pm 0.69}$	100.0	1.8	0.88
FreshPER (Ma et al., 2026)	$21.47_{\pm 1.01}$	$7.64_{\pm 0.63}$	100.0	1.5	0.84
Count-ratio correction	$25.54_{\pm 0.72}$	$9.87_{\pm 0.49}$	94.3	7.8	0.59
Age-only weighting ($\tau = 500$)	$20.38_{\pm 1.07}$	$7.08_{\pm 0.66}$	100.0	0.6	0.90
One component changed from ESR					
Lower probe reservation ($\alpha = 0.05$)	$25.69_{\pm 0.78}$	$10.02_{\pm 0.53}$	89.7	6.0	0.57
Weaker ridge ($\lambda = 10^{-4}$)	$\underline{25.96_{\pm 0.76}}$	$\underline{10.17_{\pm 0.51}}$	94.4	9.1	0.54
Without trajectory cap	$25.32_{\pm 0.94}$	$9.75_{\pm 0.64}$	94.5	0.0†	0.44
Without two-outcome gate‡	$25.61_{\pm 0.84}$	$9.92_{\pm 0.57}$	98.2	10.4	0.50
ESR (full)	$\mathbf{26.21}_{\pm 0.68}$	$\mathbf{10.36}_{\pm 0.46}$	94.6	6.9	0.63

† Cap frequency is zero by definition when capping is disabled. ‡ Missing source outcomes are extrapolated; invalid records and wholly unprobed tools remain excluded. Coverage, cap frequency, and nESS summarize replay behavior; rankings use AUC. Defaults: $\alpha = 0.1$, $\lambda = 10^{-3}$, $c = 10$.

C.5.2 Frozen-policy correction

Table 10: **Frozen-policy correction.** $0.7 \to 0.5$, with policy ratios equal to one; three-seed mean and s.d. Short/long strata use 1–4/≥ 5 affected executions. Exact ratios are uncapped; fresh-reference comparisons quantify Monte Carlo variability. Bold/underline rank numeric means within each stratum, including these diagnostic references. Metric definitions: Appendix B.7.

	Log-ratio	Gradient contribution	
Method	MAE ↓	Rel. error ↓	Cosine ↑
A. Short trajectories: 1–4 affected executions			
Policy-only replay	$0.389_{\pm 0.006}$	$0.44_{\pm 0.06}$	$0.84_{\pm 0.04}$
Count-ratio correction	$0.046_{\pm 0.012}$	$0.20_{\pm 0.03}$	$0.971_{\pm 0.010}$
ESR (full)	$\underline{0.031_{\pm 0.009}}$	$0.17_{\pm 0.03}$	$0.982_{\pm 0.007}$
Exact ratio (uncapped)	$\mathbf{0.000}_{\pm 0.000}$	$\underline{0.15_{\pm 0.03}}$	$\underline{0.986_{\pm 0.006}}$
Independent fresh reference	n/a	$\mathbf{0.12}_{\pm 0.02}$	$\mathbf{0.992}_{\pm 0.004}$
B. Long trajectories: at least 5 affected executions			
Policy-only replay	$0.392_{\pm 0.010}$	$1.02_{\pm 0.14}$	$0.31_{\pm 0.11}$
Count-ratio correction	$0.052_{\pm 0.015}$	$0.38_{\pm 0.06}$	$0.889_{\pm 0.032}$
ESR (full)	$\underline{0.036_{\pm 0.011}}$	$\underline{0.30_{\pm 0.05}}$	$\underline{0.931_{\pm 0.023}}$
Exact ratio (uncapped)	$\mathbf{0.000}_{\pm 0.000}$	$0.34_{\pm 0.06}$	$0.912_{\pm 0.029}$
Independent fresh reference	n/a	$\mathbf{0.19}_{\pm 0.03}$	$\mathbf{0.969}_{\pm 0.013}$

Sample counts (seeds 0/1/2): source trajectories 4,096 / 4,096 / 4,096; valid probe pairs 7,814 / 7,692 / 7,887; trajectories per target-reference batch 8,192 / 8,192 / 8,192. Each seed uses three independent target-reference batches; undefined metrics are marked n/a.

C.5.3 STATIC CONTROLS AND RESOURCE COSTS

The static control exactly matches policy-only replay. Under the null switch, the one-sided lower bounds of $-0.73/-0.55$ points satisfy the -1-point non-inferiority margin for both splits (Table 11). Full ESR uses 109.6K probe/restoration calls and 72.4 wall-hours, versus 54.7K and 67.2 hours at $\alpha = 0.05$ (Table 12). Halving the probe reservation lowers mean Test-N/Test-C AUC by 0.52/0.34 points, trading adaptation performance for lower probe and wall-clock costs.

Table 11: **Stability without a transition change.** Reliability 0.7, 10^6 calls; terminal TGC (%), three-seed mean and sample s.d. Difference rows give ESR minus policy-only with a one-sided 95% interval $[L, +\infty)$. Non-inferiority requires $L > -1$ for each split/setting. The null switch clears fits and retains ordinary probe costs. Differences use unrounded seed-level scores.

Setting / method	Test-N	Test-C
A. Static environment; no version signal		
Policy-only replay	$\mathbf{33.13}_{\pm 0.82}$	$\mathbf{14.27}_{\pm 0.57}$
ESR (full)	$\mathbf{33.13}_{\pm 0.82}$	$\mathbf{14.27}_{\pm 0.57}$
ESR − policy-only [95% interval]	$0.00\ [0.00, +\infty)$	$0.00\ [0.00, +\infty)$
B. Null switch at 40%; transition law unchanged		
Policy-only replay	$\mathbf{33.13}_{\pm 0.82}$	$\mathbf{14.27}_{\pm 0.57}$
ESR (full)	$\underline{32.79}_{\pm 0.82}$	$\underline{14.05}_{\pm 0.60}$
ESR − policy-only [95% interval]	$-0.35\ [-0.73, +\infty)$	$-0.22\ [-0.55, +\infty)$

Table 12: **Cost and uncertainty.** Main $0.7 \to 0.5$ adaptation protocol. A: actual training costs, three-seed means (K: 10^3 calls). B: individual seed AUCs. C: paired AUC differences with 95% bootstrap intervals. The lower-α row reuses its ablation; B/C use full ESR.

A. Measured resource use

	Charged executions (K)				Nested	Time (h)	
Method	Fresh	Probes	Restore	Total	API (K)	GPU	Wall
Fresh-only	996.4	0.0	0.0	996.4	3,928	418.6	62.3
Policy-only replay	995.8	0.0	0.0	995.8	3,914	432.7	64.9
Count-ratio correction	885.7	57.0	52.4	995.1	3,996	445.1	70.8
ESR ($\alpha = 0.05$)	939.1	28.3	26.4	993.8	3,956	439.2	67.2
ESR ($\alpha = 0.10$)	884.6	57.5	52.1	994.2	4,012	451.3	72.4

B. Seed-level AUCs

	Test-N			Test-C		
Method	Seed 0	Seed 1	Seed 2	Seed 0	Seed 1	Seed 2
Fresh-only	20.77	22.46	22.29	7.22	8.34	8.20
Policy-only replay	17.87	19.90	19.71	5.78	7.00	6.96
Count-ratio correction	24.71	26.00	25.91	9.30	10.16	10.15
ESR (full)	**25.43**	**26.70**	**26.50**	**9.83**	**10.65**	**10.60**

C. Paired differences

Contrast	Test-N Δ [95% CI]	Test-C Δ [95% CI]
ESR − fresh-only	+4.37 [3.68, 5.07]	+2.44 [1.93, 2.94]
ESR − policy-only	+7.05 [6.23, 7.88]	+3.78 [3.18, 4.39]
ESR − count-ratio	+0.67 [−0.08, 1.41]	+0.49 [−0.05, 1.03]

Charged categories are disjoint, include failed calls and both probe sides, and sum to at most 10^6/seed. Nested APIs are separate telemetry. GPU-hours include all training computation; wall time includes restoration. Evaluation and diagnostics are accounted for separately.

C.6 GRADIENT CORRECTION OVER TRAINING

In the frozen-policy diagnostic (Table 10), both count-ratio correction and ESR improve gradient alignment in the short and long strata. With policy ratios fixed at one, these improvements reflect accumulated environment mismatch without policy staleness.

Figure 5 follows gradient direction and scale through training using the protocol and coordinates in Appendix B.8. Immediately after the switch, ESR raises mean cosine from 0.26 to 0.89 and reduces median absolute scale mismatch $|s_m|$ from 0.36 to 0.06; uncapped exact ratios reach mean cosine 0.94. By the final checkpoint, mean cosines reach 0.92 for policy-only replay and 0.97 for ESR, while median $|s_m|$ is 0.14 and 0.06, respectively. Cosine is invariant to positive scalar rescaling, so the early gain reflects a changed update direction. The narrowing gap as stale records are replaced is consistent with correction targeting environment-stale replay. These diagnostics measure the joint effect of fitting, gating, and capping.

C.7 NUMERICAL SENSITIVITY TO WITHIN-LABEL CHANGES

The model and grid are specified in Appendix B.10. Along $\kappa = 0$, mean gain increases from $+0.068$ at $\delta = 0.06$ to $+0.276$ at $\delta = 0.24$. Increasing κ narrows the positive-gain region: for $\delta \in \{0.18, 0.20, 0.22, 0.24\}$, linear interpolation between adjacent grid means places the zero crossings at $\kappa \approx 0.156$–0.171. At $(\delta, \kappa) = (0.24, 0.24)$, mean gain is -0.086; the absolute ESR and policy-only cosines are 0.245 and 0.331, respectively. These score patterns illustrate the distinction between correctable frequency drift and unmodeled changes within a tool–outcome cell. The corresponding identifiability limitation is established by Proposition 2.

D PROOFS

This appendix proves the results in Section 4.4, from population identification and gradient recovery to the limits imposed by feature coarsening and missing support. Random variables and kernels are defined on standard Borel spaces. Vector norms are Euclidean, matrix inequalities use the positive-semidefinite order, and density identities hold almost surely under the stated reference law. Rewards, environment versions, scoring policies, and baselines are fixed as specified below.

D.1 POPULATION ASSUMPTIONS

Fix the policy parameter θ, source version v, behavior policy β, and reward baseline $b \in [0,1]$ independently of the trajectories being analyzed. Initial laws, stopping rules, and the reward $R \in [0,1]$ are common across versions, and $T \le H$. The behavior policy covers the current policy's actions. The action support is parameter-independent; in a convex neighborhood of θ, the policy is continuously differentiable and each complete-turn log-policy gradient has norm at most G. Terminated trajectories are padded with absorbing transitions and zero scores. These conditions justify differentiation under the trajectory integral. For identification, each affected tool has positive probe-anchor probability, both source labels have positive probability, and retained probe pairs have the conditional marginals prescribed in Section 4.2. These assumptions concern the population distributions; empirical counts provide the separate evidence checks used by the algorithm.

The analysis applies least-squares density-ratio estimation and score-function differentiation (Kanamori et al., 2009; Sutton et al., 1999) to paired tool executions, characterizes information lost through the tool–outcome encoding, and derives sharp gradient-bias bounds. Exact recovery uses population correction and independent fixed-source sampling. For the practical update in Equation (9), Appendix D.5 bounds the effects of estimation, ridge, rejection, and capping, and explains the dependence introduced by selecting θ from a realized finite buffer.

D.2 POPULATION LEAST-CHANGE CHARACTERIZATION

Let $Q^+ \ll Q^-$, $\rho = dQ^+/dQ^- \in L^2(Q^-)$, and let ϕ have finite second moments. Define $\mu = \mathbb{E}_{Q^-}\phi$, $\psi = \phi - \mu$, $C = \mathbb{E}_{Q^-}[\psi\psi^\top]$, and $\Delta = \mathbb{E}_{Q^+}\phi - \mu$. The minimum of $\mathbb{E}_{Q^-}(r-1)^2$ over square-integrable signed functions satisfying $\mathbb{E}_{Q^-} r = 1$ and $\mathbb{E}_{Q^-}[r\phi] = \mathbb{E}_{Q^+}\phi$ is attained uniquely, up to null sets, by

$$r_\phi = 1 + \psi^\top C^\dagger \Delta, \qquad \|\rho - r_\phi\|^2_{L^2(Q^-)} = \chi^2(Q^+\|Q^-) - \Delta^\top C^\dagger \Delta. \tag{20}$$

Here $C^\dagger$ is the Moore–Penrose pseudoinverse and $\chi^2(Q^+\|Q^-) = \mathbb{E}_{Q^-}(\rho-1)^2$.

Proof. Since $\Delta = \mathbb{E}_{Q^-}[\psi(\rho-1)]$, every $v_0 \in \ker C$ satisfies $v_0^\top \psi = 0$ almost surely and $v_0^\top \Delta = 0$. Thus $\Delta \in \mathrm{range}(C)$. Taking $w_0 = C^\dagger \Delta$ gives $Cw_0 = \Delta$, so r_ϕ satisfies both constraints. For another feasible r, the difference $e = r - r_\phi$ has zero mean and is orthogonal to every centered feature. Consequently $\|r-1\|^2 = \|r_\phi - 1\|^2 + \|e\|^2$, proving minimality and uniqueness. Taking $r = \rho$ gives the stated error identity because $\|r_\phi - 1\|^2 = \Delta^\top C^\dagger \Delta$. □

Under Equation (2), ρ is constant in each tool–outcome cell, so $\rho - 1$ lies in the centered one-hot span on probe support and $r_\phi = \rho$ there. General within-cell dynamics changes produce an approximation residual. With population ridge coefficient $w_\lambda = (C + \lambda I)^{-1}\Delta$, the residual moment is $Cw_\lambda - \Delta = -\lambda w_\lambda$. If $C = U\,\mathrm{diag}(\sigma_j)U^\top$, each coefficient along $U_{:j}$ is $(U_{:j}^\top \Delta)/(\sigma_j + \lambda)$. When the two environment laws coincide, $\Delta = 0$ and the population correction is one; finite paired draws can still differ. Known identical versions use their exact unit factor.

Held-out estimation risk. For an independently scored paired sample, the diagnostic $\widehat{\mathcal{R}}(r) = (2m_{\mathrm{val}})^{-1}\sum_j r(u_j, z_j^-)^2 - m_{\mathrm{val}}^{-1}\sum_j r(u_j, z_j^+)$ estimates $\mathcal{R}(r) = \frac{1}{2}\|r-\rho\|^2_{L^2(Q^-)} - \frac{1}{2}\|\rho\|^2_{L^2(Q^-)}$. It differs from *half* the squared ratio error by a constant. Finite-probe accuracy depends on the source cell probabilities and the allocated sample counts.

D.3 IDENTIFICATION FROM PAIRED PROBES AND EXACT FEATURE RESIDUALS

For one source version and scoring fold, let $Q^- = \nu P_v$ and $Q^+ = \nu P_{v_\star}$ denote the two joint anchor–outcome laws. Let C_j be the jth tool–outcome cell, $\phi = e_j$ on C_j, $\mu = \mathbb{E}_{Q^-}\phi$, $\mu^+ = \mathbb{E}_{Q^+}\phi$, $\psi = \phi - \mu$, $C = \mathbb{E}_{Q^-}\psi\psi^\top$, and $\Delta = \mu^+ - \mu$. Write $\mathrm{TV}(P, Q) = \sup_A |P(A) - Q(A)|$ and $\chi^2(Q\|P) = \mathbb{E}_P[(dQ/dP - 1)^2]$ when $Q \ll P$.

Theorem 2 (Paired identification and exact feature residual)**.** *Suppose $Q^+ \ll Q^-$, $\rho = dQ^+/dQ^- \in L^2(Q^-)$, and $\mu_j > 0$ for every cell. The unregularized population correction is*

$$r_\phi = 1 + \psi^\top C^\dagger \Delta = \mathbb{E}_{Q^-}[\rho \mid \phi], \qquad r_\phi\big|_{C_j} = \frac{\mu_j^+}{\mu_j}. \tag{21}$$

It is the unique minimum of $\mathbb{E}_{Q^-}(r-1)^2$ over signed square-integrable r satisfying $\mathbb{E}_{Q^-} r = 1$ and $\mathbb{E}_{Q^-}[r\phi] = \mu^+$. If $Q_j^\pm = Q^\pm(\cdot \mid C_j)$, with zero-target-mass terms omitted, its exact residuals are

$$\|\rho - r_\phi\|^2_{L^2(Q^-)} = \chi^2(Q^+\|Q^-) - \Delta^\top C^\dagger \Delta = \sum_{j:\mu_j^+>0} \frac{(\mu_j^+)^2}{\mu_j} \chi^2(Q_j^+\|Q_j^-),$$

$$\sup_{0\le f\le 1} |\mathbb{E}_{Q^+} f - \mathbb{E}_{Q^-}[r_\phi f]| = \mathrm{TV}(Q^+, r_\phi Q^-) = \sum_{j:\mu_j^+>0} \mu_j^+ \,\mathrm{TV}(Q_j^+, Q_j^-) \le \tfrac{1}{2}\|\rho - r_\phi\|_{L^2(Q^-)}. \tag{22}$$

The constant $1/2$ is sharp. Under Equation (2)*, the within-cell laws agree, so $r_\phi = \rho = p_{v_\star}(y \mid k)/p_v(y \mid k)$. The identified ratio applies at every replay anchor of a covered tool, including anchors outside the probe support, because the mixing probabilities depend only on tool identity.*

To prove Theorem 2, we first identify the accepted probe laws and then derive the feature residuals.

The law of an accepted pair. Let D indicate that a probe pair is retained. Let ν denote the law of its restored anchor U conditional on $D = 1$. The accepted-pair contract is

$$\begin{aligned} &\mathcal{L}(Z^- \mid U = u, D = 1) = P_v(\cdot \mid u), \\ &\mathcal{L}(Z^+ \mid U = u, D = 1) = P_{v_\star}(\cdot \mid u), \qquad \nu\text{-almost every } u. \end{aligned} \tag{23}$$

For example, selecting an accessible anchor before execution and retaining every execution governed by the stated wrapper model satisfies this contract. Outcome-dependent rejection can alter the required conditional marginals. Independent branch randomness, as used by the algorithm, is sufficient for the prescribed joint execution scheme; the population identification argument only needs the two conditional marginals in Equation (23).

Lemma 1 (Cancellation of the common anchor law)**.** *Assume Equation* (23)*. Suppose $P_{v_\star}(\cdot \mid u) \ll P_v(\cdot \mid u)$ for ν-almost every u, with a jointly measurable conditional density $\rho(u, z)$. Define $Q^- = \nu P_v$ and $Q^+ = \nu P_{v_\star}$. Then $Q^+ \ll Q^-$ and $dQ^+/dQ^- = \rho$. Under Equation* (2)*, this density is $p_{v_\star}(y \mid k)/p_v(y \mid k)$ on the tool–outcome cell (k, y) whenever $p_v(y \mid k) > 0$.*

Proof. For an arbitrary measurable set A in the anchor–outcome product space, the definition of a conditional density and Tonelli's theorem give

$$\begin{aligned} Q^+(A) &= \int \nu(du) \int \mathbf{1}_A(u, z) P_{v_\star}(dz \mid u) \\ &= \int \nu(du) \int \mathbf{1}_A(u, z)\rho(u, z) P_v(dz \mid u) \\ &= \int_A \rho(u, z) Q^-(du, dz). \end{aligned} \tag{24}$$

In particular, $Q^-(A) = 0$ implies $Q^+(A) = 0$, proving absolute continuity. Equation (24) for every measurable A proves the asserted Radon–Nikodym identity. To see uniqueness directly, if another density $\widetilde{\rho}$ gives the same integrals, then for $A_\varepsilon = \{\rho \ge \widetilde{\rho} + \varepsilon\}$ and every $\varepsilon > 0$,

$$0 = \int_{A_\varepsilon} (\rho - \widetilde{\rho})\, dQ^- \ge \varepsilon Q^-(A_\varepsilon) \ge 0. \tag{25}$$

Taking a countable union over positive rational ε yields $Q^-(\rho > \widetilde{\rho}) = 0$; interchanging the densities proves equality almost surely. Thus the common anchor distribution contributes no additional visitation ratio.

The disjoint-support assumption gives the conditional density directly from label probabilities. For a fixed anchor u with tool k, let $B_y(u)$ be the measurable region identified by label y. The kernel $G_y(\cdot \mid u)$ gives this region probability one, and gives the other label's region probability zero. For any measurable outcome set B,

$$\begin{aligned}
&\int_B \frac{p_{v_\star}(y(z) \mid k)}{p_v(y(z) \mid k)} P_v(dz \mid u) \\
&\quad = \sum_{y=0}^{1} p_v(y \mid k) \int_{B \cap B_y(u)} \frac{p_{v_\star}(y \mid k)}{p_v(y \mid k)} G_y(dz \mid u) \\
&\quad = \sum_{y=0}^{1} p_{v_\star}(y \mid k) G_y(B \mid u) = P_{v_\star}(B \mid u).
\end{aligned} \tag{26}$$

This identifies the density, including the case of a zero target probability. Domination requires positive source probability for each outcome with positive target probability.

Write $\alpha_k = \nu\{u : k(u) = k\}$ and $C_{k,y} = \{(u, z) : k(u) = k, y(z) = y\}$. Whenever $\alpha_k > 0$, the source and target cell probabilities obey

$$\begin{aligned}
&Q^-(C_{k,y}) = \alpha_k p_v(y \mid k), \qquad && Q^+(C_{k,y}) = \alpha_k p_{v_\star}(y \mid k), \\
&\frac{Q^+(C_{k,y})}{Q^-(C_{k,y})} = \frac{\alpha_k p_{v_\star}(y \mid k)}{\alpha_k p_v(y \mid k)} = \frac{p_{v_\star}(y \mid k)}{p_v(y \mid k)}.
\end{aligned} \tag{27}$$

Hence positive probe mass for a tool identifies its two correction factors. The state-independent mixing probabilities and version-independent $G_y(\cdot \mid u)$ extend their validity to other restored states of that tool, including states outside the support of ν. A tool with $\alpha_k = 0$ remains unidentified by Equation (27). □

Lemma 2 (Least-change correction with a singular covariance)**.** *Let $Q^+ \ll Q^-$ with $\rho = dQ^+/dQ^- \in L^2(Q^-)$. For any square-integrable feature vector $\phi \in \mathbb{R}^d$, set $\mu = \mathbb{E}_{Q^-}\phi$, $\psi = \phi - \mu$, $C = \mathbb{E}_{Q^-}[\psi\psi^\top]$, and $\Delta = \mathbb{E}_{Q^+}\phi - \mu$. Then $\Delta \in \mathrm{range}(C)$ and*

$$\begin{aligned}
r_\phi &= 1 + \psi^\top C^\dagger \Delta, \\
\|\rho - r_\phi\|^2_{L^2(Q^-)} &= \chi^2(Q^+ \| Q^-) - \Delta^\top C^\dagger \Delta.
\end{aligned} \tag{28}$$

Moreover, r_ϕ uniquely minimizes $\mathbb{E}_{Q^-}(r-1)^2$ over signed $r \in L^2(Q^-)$ satisfying $\mathbb{E}_{Q^-} r = 1$ and $\mathbb{E}_{Q^-}[r\phi] = \mathbb{E}_{Q^+}\phi$.

Proof. All feature moments under Q^+ exist: by Cauchy–Schwarz, $\mathbb{E}_{Q^-}[|\rho\phi_j|] \le \|\rho\|_2 \|\phi_j\|_2 < \infty$ for each coordinate. Since $\mathbb{E}_{Q^-}\rho = 1$ and $\mathbb{E}_{Q^-}\psi = 0$,

$$\Delta = \mathbb{E}_{Q^-}[(\rho - 1)\phi] = \mathbb{E}_{Q^-}[(\rho - 1)\psi]. \tag{29}$$

If $a \in \ker C$, the following chain proves that $a^\top \Delta = 0$:

$$\begin{aligned}
|a^\top \Delta| &= \left|\mathbb{E}_{Q^-}[(a^\top \psi)(\rho - 1)]\right| \\
&\le \mathbb{E}_{Q^-}[|a^\top \psi|\,|\rho - 1|] \\
&\le \left(\mathbb{E}_{Q^-}(a^\top \psi)^2\right)^{1/2} \left(\mathbb{E}_{Q^-}(\rho - 1)^2\right)^{1/2} \\
&= \left(a^\top C a\right)^{1/2} \chi^2(Q^+ \| Q^-)^{1/2} = 0.
\end{aligned} \tag{30}$$

Because C is symmetric, $\mathrm{range}(C) = (\ker C)^\perp$. Thus Δ lies in this range, even when C is singular. Set $w_0 = C^\dagger \Delta$; the Moore–Penrose identities give $C w_0 = \Delta$. The proposed function satisfies the constraints exactly:

$$\begin{aligned}
\mathbb{E}_{Q^-} r_\phi &= 1 + \mathbb{E}_{Q^-}[\psi]^\top w_0 = 1, \\
\mathbb{E}_{Q^-}[r_\phi \phi] &= \mu + \mathbb{E}_{Q^-}[(\psi^\top w_0)(\psi + \mu)] = \mu + C w_0 = \mu + \Delta = \mathbb{E}_{Q^+}\phi.
\end{aligned} \tag{31}$$

For any other feasible r, let $e = r - r_\phi$. Subtracting its moment constraints from Equation (31) gives $\mathbb{E}_{Q^-} e = 0$ and $\mathbb{E}_{Q^-}[e\psi] = 0$. Therefore the cross term vanishes, and the exact Pythagorean decomposition yields

$$\begin{aligned}\mathbb{E}_{Q^-}(r-1)^2 &= \mathbb{E}_{Q^-}\left[e + (r_\phi - 1)\right]^2 \\ &= \mathbb{E}_{Q^-} e^2 + 2w_0^\top \mathbb{E}_{Q^-}[e\psi] + \mathbb{E}_{Q^-}(r_\phi - 1)^2 \\ &= \mathbb{E}_{Q^-} e^2 + \mathbb{E}_{Q^-}(r_\phi - 1)^2 \\ &\geq \mathbb{E}_{Q^-}(r_\phi - 1)^2 \geq 0.\end{aligned} \tag{32}$$

Equality in the first inequality holds if and only if $\mathbb{E}_{Q^-} e^2 = 0$, proving uniqueness up to source-null sets. The true density ρ is feasible. Applying Equation (32) with $r = \rho$, and computing the correction norm, gives

$$\begin{aligned}\mathbb{E}_{Q^-}(r_\phi - 1)^2 &= w_0^\top C w_0 = \Delta^\top C^\dagger \Delta, \\ \chi^2(Q^+ \| Q^-) &= \|\rho - r_\phi\|^2_{L^2(Q^-)} + \Delta^\top C^\dagger \Delta \geq \Delta^\top C^\dagger \Delta \geq 0.\end{aligned} \tag{33}$$

This proves Equation (28) without an invertibility assumption.

The same function is the L^2 projection of ρ onto the affine centered feature family. Indeed, with $r_w = 1 + \psi^\top w$, expansion around w_0 gives

$$\begin{aligned}\frac{1}{2} w^\top C w - \Delta^\top w &= \frac{1}{2}(w - w_0)^\top C (w - w_0) - \frac{1}{2}\Delta^\top C^\dagger \Delta \\ &\geq -\frac{1}{2}\Delta^\top C^\dagger \Delta, \\ \|\rho - r_w\|^2_{L^2(Q^-)} &= \|\rho - r_\phi\|^2_{L^2(Q^-)} + (w - w_0)^\top C (w - w_0) \\ &\geq \|\rho - r_\phi\|^2_{L^2(Q^-)}.\end{aligned} \tag{34}$$

Any coefficient minimizer differs from w_0 by an element of $\ker C$, which changes the fitted function only on a source-null set. □

Lemma 3 (Exact residual of one-hot coarsening)**.** *Under Lemma 2, let $\phi = e_j$ on a measurable partition cell C_j. Write $\mu_j = Q^-(C_j)$, $\mu_j^+ = Q^+(C_j)$, and restrict to cells with $\mu_j > 0$; cells with zero source mass also have zero target mass and are irrelevant to the almost-sure statements. For $\mu_j^+ > 0$, let $Q_j^\pm = Q^\pm(\cdot \mid C_j)$. Then*

$$\begin{aligned}r_\phi\big|_{C_j} &= \frac{\mu_j^+}{\mu_j}, & r_\phi &= \mathbb{E}_{Q^-}[\rho \mid \phi], \\ \|\rho - r_\phi\|^2_{L^2(Q^-)} &= \sum_{j:\mu_j^+ > 0} \frac{(\mu_j^+)^2}{\mu_j} \chi^2(Q_j^+ \| Q_j^-), \\ \mathrm{TV}(Q^+, r_\phi Q^-) &= \sum_{j:\mu_j^+ > 0} \mu_j^+ \,\mathrm{TV}(Q_j^+, Q_j^-).\end{aligned} \tag{35}$$

Here $\mathrm{TV}(P, Q) = \sup_A |P(A) - Q(A)|$ for probability laws. In particular, the fitted law equals the target law if and only if the two within-cell conditional laws agree in every positive-target cell.

Proof. Conditional expectation on a positive-source cell gives

$$\mathbb{E}_{Q^-}[\rho \mid C_j] = \frac{1}{\mu_j}\int_{C_j} \rho \, dQ^- = \frac{Q^+(C_j)}{\mu_j} = \frac{\mu_j^+}{\mu_j}. \tag{36}$$

For an explicit connection to the covariance solve, let $v_j = \mu_j^+ / \mu_j$ on the active coordinates. Since $\mu^\top v = 1$, the one-hot covariance and shift satisfy

$$\begin{aligned}C &= \mathrm{diag}(\mu) - \mu\mu^\top, & \Delta &= \mu^+ - \mu, \\ Cv &= \mathrm{diag}(\mu) v - \mu(\mu^\top v) = \mu^+ - \mu = \Delta, \\ 1 + (e_j - \mu)^\top v &= 1 + v_j - \mu^\top v = v_j.\end{aligned} \tag{37}$$

Lemma 2 then implies that this fitted function equals r_ϕ. In particular it is nonnegative and normalized, so $r_\phi Q^-$ is a probability law. If $\mu_j^+ = 0$, nonnegativity of ρ and $\int_{C_j} \rho \, dQ^- = 0$ imply $\rho = r_\phi = 0$ on that cell, almost surely.

For $\mu_j^+ > 0$, the conditional target law is dominated by the conditional source law. Write $\eta_j = dQ_j^+/dQ_j^-$. For every measurable $A \subseteq C_j$,

$$\begin{aligned} \int_A \rho \, dQ^- = Q^+(A) = \mu_j^+ Q_j^+(A) = \mu_j^+ \int_A \eta_j \, dQ_j^- \\ = \int_A \frac{\mu_j^+}{\mu_j} \eta_j \, dQ^-. \end{aligned} \tag{38}$$

Uniqueness of densities therefore yields $\rho = (\mu_j^+/\mu_j)\eta_j$ on C_j. Integrating its centered square proves the exact residual formula:

$$\begin{aligned} \|\rho - r_\phi\|_{L^2(Q^-)}^2 &= \sum_{j:\mu_j^+>0} \int_{C_j} \left(\rho - \frac{\mu_j^+}{\mu_j}\right)^2 dQ^- \\ &= \sum_{j:\mu_j^+>0} \mu_j \left(\frac{\mu_j^+}{\mu_j}\right)^2 \int (\eta_j - 1)^2 dQ_j^- \\ &= \sum_{j:\mu_j^+>0} \frac{(\mu_j^+)^2}{\mu_j} \chi^2(Q_j^+ \| Q_j^-). \end{aligned} \tag{39}$$

The conditional-mean identity also gives $\mathbb{E}_{Q^-}[\rho - r_\phi \mid \phi] = 0$. Expanding the square of $\rho - 1 = (\rho - r_\phi) + (r_\phi - 1)$, its cross term is zero, so the total divergence decomposes exactly as

$$\begin{aligned} \chi^2(Q^+ \| Q^-) &= \mathbb{E}_{Q^-}(r_\phi - 1)^2 + \mathbb{E}_{Q^-}(\rho - r_\phi)^2 \\ &= \sum_{j:\mu_j>0} \frac{(\mu_j^+ - \mu_j)^2}{\mu_j} + \sum_{j:\mu_j^+>0} \frac{(\mu_j^+)^2}{\mu_j} \chi^2(Q_j^+ \| Q_j^-). \end{aligned} \tag{40}$$

Thus the first term is precisely the part captured by the cell frequencies, and the second is the remaining within-cell error.

To establish the total-variation identity and its constant, set $\xi = \rho - r_\phi$. Both densities integrate to one, so $\mathbb{E}_{Q^-}\xi = 0$. Writing $\xi_+ = \max(\xi, 0)$ and $\xi_- = \max(-\xi, 0)$ gives

$$\begin{aligned} \mathbb{E}_{Q^-}\xi_+ = \mathbb{E}_{Q^-}\xi_- = \tfrac{1}{2}\mathbb{E}_{Q^-}|\xi|, \\ -\mathbb{E}_{Q^-}\xi_- \le \int_A \xi \, dQ^- \le \mathbb{E}_{Q^-}\xi_+ \qquad \text{for every measurable } A, \\ \sup_A \left| \int_A \xi \, dQ^- \right| = \int_{\{\xi>0\}} \xi \, dQ^- = \tfrac{1}{2}\mathbb{E}_{Q^-}|\xi|. \end{aligned} \tag{41}$$

The admissible set $\{\xi > 0\}$ attains the last equality. Applying this identity first to the joint laws and then to each conditional pair yields

$$\begin{aligned} 2\,\mathrm{TV}(Q^+, r_\phi Q^-) &= \sum_{j:\mu_j^+>0} \mu_j \int \left| \frac{\mu_j^+}{\mu_j}(\eta_j - 1) \right| dQ_j^- \\ &= \sum_{j:\mu_j^+>0} \mu_j^+ \int |\eta_j - 1| dQ_j^- \\ &= 2 \sum_{j:\mu_j^+>0} \mu_j^+ \,\mathrm{TV}(Q_j^+, Q_j^-). \end{aligned} \tag{42}$$

Every summand is nonnegative, and every displayed coefficient is positive. The residual is zero exactly when each corresponding conditional law agrees, as stated in the lemma. □

Bounded tests and optimal constants. For any measurable f with $0 \le f \le 1$, the fact that $\xi = \rho - r_\phi$ has zero mean permits centering at $1/2$. Consequently,

$$\begin{aligned}\left|\mathbb{E}_{Q^+} f - \mathbb{E}_{Q^-}[r_\phi f]\right| &= \left|\mathbb{E}_{Q^-}[\xi(f - \tfrac{1}{2})]\right| \\ &\le \mathbb{E}_{Q^-}[|\xi|\,|f - \tfrac{1}{2}|] \le \tfrac{1}{2}\mathbb{E}_{Q^-}|\xi| \\ &= \mathrm{TV}(Q^+, r_\phi Q^-) = \sum_{j:\mu_j^+>0} \mu_j^+ \,\mathrm{TV}(Q_j^+, Q_j^-) \\ &\le \tfrac{1}{2}\left(\mathbb{E}_{Q^-}\xi^2\right)^{1/2}\left(\mathbb{E}_{Q^-}1^2\right)^{1/2} \\ &= \tfrac{1}{2}\left[\sum_{j:\mu_j^+>0} \frac{(\mu_j^+)^2}{\mu_j}\chi^2(Q_j^+\|Q_j^-)\right]^{1/2} \\ &= \tfrac{1}{2}\left[\chi^2(Q^+\|Q^-) - \Delta^\top C^\dagger \Delta\right]^{1/2}.\end{aligned} \tag{43}$$

The supremum over these tests equals the total variation exactly: $f = \mathbf{1}_{\{\xi>0\}}$ attains it by Equation (41). The last bound can also be combined with the elementary bound $\mathrm{TV}(Q^+, r_\phi Q^-) \le 1$.

The factor $1/2$ in the L^2 bound is sharp. To exhibit an equality case, take any positive probability vector $(\mu_j)_{j=1}^d$ on the cells and split each cell into two atoms (j,σ) with $\sigma \in \{-1,1\}$. For $a \in (0,1]$, define

$$\begin{aligned} Q^-(j,\sigma) = \frac{\mu_j}{2}, \quad Q^+(j,\sigma) = \frac{\mu_j(1+\sigma a)}{2}, \\ \mu_j^+ = \mu_j, \qquad r_\phi(j,\sigma) = 1, \\ \rho(j,\sigma) - r_\phi(j,\sigma) = \sigma a. \end{aligned} \tag{44}$$

These are probability laws and the target is dominated by the source. For the test $f(j,\sigma) = \mathbf{1}_{\{\sigma=1\}}$,

$$\begin{aligned} \left|\mathbb{E}_{Q^+} f - \mathbb{E}_{Q^-}[r_\phi f]\right| &= \frac{1+a}{2} - \frac{1}{2} = \frac{a}{2}, \\ \mathrm{TV}(Q^+, r_\phi Q^-) &= \tfrac{1}{2}\sum_{j,\sigma}\frac{\mu_j}{2}|\sigma a| = \frac{a}{2} = \tfrac{1}{2}\|\rho - r_\phi\|_{L^2(Q^-)}. \end{aligned} \tag{45}$$

This example illustrates the general coarsening result through a within-cell change.

Finally, under the homogeneous reliability model and for $\alpha_k > 0$, the conditional law in every cell of positive target mass is identical across versions. More explicitly, for every measurable set A,

$$\begin{aligned} Q^-\big(A \mid C_{k,y}\big) &= \int \nu(du \mid k(U) = k) \int \mathbf{1}_A(u,z) G_y(dz \mid u) \\ &= Q^+\big(A \mid C_{k,y}\big). \end{aligned} \tag{46}$$

The kernels' disjoint, label-identified supports justify conditioning on y, and their mixing probabilities are constant across anchors of a given tool. Equation (46) makes every conditional divergence in Lemma 3 zero. Lemma 1 and Equation (27) identify the remaining cell values as the exact environment ratios. Lemma 2 handles singular covariance and gives the Pythagorean identity. Combining these results with Equation (43) proves Theorem 2.

D.4 Trajectory change of measure and gradient consistency

Let $P_0 = p_{\beta,v}$, $P = p_{\theta,v}$, and $Q = p_{\theta,v_\star}$. Define the policy ratio $A = \prod_{t<T} \pi_\theta(a_t \mid h_t)/\beta(a_t \mid h_t)$, environment ratio $L = \prod_{t<T} \rho_v^E(u_t, z_t)$, complete weight $W = AL$, and trajectory score $S = \sum_{t<T} \nabla \log \pi_\theta(a_t \mid h_t)$. On source records with $A = 0$, set $S = 0$ and their weighted contributions to zero; the score is only needed on the current policy's support. Let $I_v = \mathbb{E}_P[SS^\top]$ and $M = \operatorname{ess\,sup}_{P+Q} \|S\|_2 \le HG$. Here S denotes the vector trajectory score written as s_θ in the main-text theorem.

Theorem 3 (Oracle gradient consistency and sharp omission bias)**.** *Under the population assumptions and the reliability model, $dP/dP_0 = A$, $dQ/dP = L$, and $dQ/dP_0 = W$. Consequently,*

$$g_\star := \nabla J_\star(\theta) = \mathbb{E}_{P_0}[W(R-b)S], \qquad g_v := \nabla J_v(\theta) = \mathbb{E}_{P_0}[A(R-b)S]. \tag{47}$$

For independent trajectories from the fixed source P_0, the empirical mean of $W(R-b)S$ converges almost surely to $g_\star$, without requiring a finite weight variance. Moreover,

$$\begin{aligned} \|g_\star - g_v\|_2 &\le \min\left\{ M\,\mathrm{TV}(P,Q),\ \tfrac{1}{2}\sqrt{\chi^2(Q\|P)\lambda_{\max}(I_v)} \right\}, \\ I_v &= \sum_{t=0}^{H-1} \mathbb{E}_P[D_t D_t^\top], \qquad \lambda_{\max}(I_v) \le G^2 \mathbb{E}_P T, \end{aligned} \tag{48}$$

where $D_t = \mathbf{1}\{t<T\}\nabla \log \pi_\theta(a_t \mid h_t)$. The constants 1 in the total-variation bound and $1/2$ in the Fisher–χ^2 bound are sharp over this model class.

We prove Theorem 3 using likelihood and score identities, followed by the gradient-bias bounds. Throughout the proof, θ, environment versions, and source policies are fixed, and both environments share the reward functional.

Lemma 4 (Stopped likelihoods and orthogonal score increments)**.** *Assume the horizon is bounded by H, the stopping rule is common across environment versions, and the policy has parameter-independent support. Suppose there is a convex neighborhood $\mathcal{U}$ of θ on which the full-turn log likelihood is continuously differentiable and its gradient has Euclidean norm at most G, uniformly over active histories and permitted actions. For either fixed environment $e \in \{v, v_\star\}$, put $P_e = p_{\theta,e}$ and define the padded score increments and their sum by*

$$\begin{aligned} D_t(\tau) &= \mathbf{1}\{t<T\}\nabla_\theta \log \pi_\theta(a_t \mid h_t), \qquad 0 \le t < H, \\ S(\tau) &= \sum_{t=0}^{H-1} D_t(\tau), \qquad \mathcal{I}_e = \mathbb{E}_{P_e}[SS^\top]. \end{aligned} \tag{49}$$

Then $\mathbb{E}_{P_e} S = 0$, differentiation of the bounded-reward objective is valid, and

$$\begin{aligned} \nabla J_e(\theta) &= \mathbb{E}_{P_e}[RS] = \mathbb{E}_{P_e}[(R-b)S] \qquad \textit{for every fixed } b \in \mathbb{R}, \\ \mathcal{I}_e &= \sum_{t=0}^{H-1} \mathbb{E}_{P_e}[D_t D_t^\top], \\ \lambda_{\max}(\mathcal{I}_e) &\le G^2 \mathbb{E}_{P_e} T \le HG^2, \qquad \|S\|_2 \le GT \le GH \quad P_e\text{-a.s.} \end{aligned} \tag{50}$$

Here b may be conditioned on if it is independent of the trajectories to which the lemma is applied.

Proof. After termination, append a common absorbing state, a deterministic dummy action, and deterministic dummy observations until time H. The appended factors equal one and their scores equal zero. This converts the stopped experiment into a fixed-length experiment without changing any trajectory probability, reward, or score. In particular, the original stopping time itself contributes no extra derivative: a policy-generated termination decision is already included in the probability of the complete sampled action. The same observation applies inside a generation turn, whose likelihood includes all sampled tokens, including its termination token.

Fix e and write $\theta_0 = \theta$. Choose $r > 0$ such that the closed ball of radius r about θ_0 lies inside $\mathcal{U}$. Parameter-independent support ensures that $P_{\vartheta,e}$ and $P_{\theta_0,e}$ have the same support for all ϑ in this ball. The fundamental theorem of calculus applied separately to every active action factor gives

$$\begin{aligned} \left| \log \frac{p_{\vartheta,e}(\tau)}{p_{\theta_0,e}(\tau)} \right| &= \left| \sum_{t<T} \int_0^1 (\vartheta - \theta_0)^\top \nabla_\xi \log \pi_\xi(a_t \mid h_t)\big|_{\xi=\theta_0+u(\vartheta-\theta_0)}\, du \right| \\ &\le \sum_{t<T} \int_0^1 \|\vartheta - \theta_0\|_2 \left\| \nabla_\xi \log \pi_\xi(a_t \mid h_t)\big|_{\xi=\theta_0+u(\vartheta-\theta_0)} \right\|_2 du \\ &\le GT\|\vartheta - \theta_0\|_2 \le GHr. \end{aligned} \tag{51}$$

The common initial law and the environment kernels cancel in this ratio. If $K_\vartheta = dP_{\vartheta,e}/dP_{\theta_0,e}$, differentiating its finite product and using Equation (51) yields the integrable envelope

$$\begin{aligned}\|\nabla_\vartheta K_\vartheta(\tau)\|_2 &= K_\vartheta(\tau)\left\|\sum_{t<T}\nabla_\vartheta \log \pi_\vartheta(a_t \mid h_t)\right\|_2 \\ &\le e^{GHr}\sum_{t<T}\|\nabla_\vartheta \log \pi_\vartheta(a_t \mid h_t)\|_2 \\ &\le GHe^{GHr}.\end{aligned} \tag{52}$$

The reward is independent of the policy parameters and lies in $[0,1]$. Dominated differentiation therefore applies both to $\mathbb{E}_{P_{\theta_0,e}}[RK_\vartheta]$ and to $\mathbb{E}_{P_{\theta_0,e}}K_\vartheta = 1$. Evaluating at $\vartheta = \theta_0$ gives

$$\begin{aligned}\nabla J_e(\theta_0) &= \int R(\tau)\nabla_\vartheta K_\vartheta(\tau)\big|_{\vartheta=\theta_0}\, dP_{\theta_0,e}(\tau) = \mathbb{E}_{P_e}[RS], \\ \mathbb{E}_{P_e}S &= \int \nabla_\vartheta K_\vartheta(\tau)\big|_{\vartheta=\theta_0}\, dP_{\theta_0,e}(\tau) = \nabla_\vartheta 1 = 0.\end{aligned} \tag{53}$$

This also proves the baseline identity. The score bound throughout the parameter neighborhood supplies the domination required for differentiation.

For the second-moment identity, let $\mathcal{F}_t$ contain the full padded interaction history immediately before action t, including the current latent state. The event $\{t < T\}$ is $\mathcal{F}_t$-measurable. Although the policy observes only h_t, its conditional action law given $\mathcal{F}_t$ remains $\pi_\theta(\cdot \mid h_t)$. The same bounded-derivative argument, now for this single conditional action distribution, gives

$$\begin{aligned}\mathbb{E}_{P_e}[D_t \mid \mathcal{F}_t] &= \mathbf{1}\{t<T\}\int \pi_\theta(da \mid h_t)\nabla_\theta \log \pi_\theta(a \mid h_t) \\ &= \mathbf{1}\{t<T\}\nabla_\theta \int \pi_\theta(da \mid h_t) = 0.\end{aligned} \tag{54}$$

The integral includes every token in the finite sampled generation. For $s < t$, D_s is $\mathcal{F}_t$-measurable and all products are integrable, so

$$\begin{aligned}\mathbb{E}_{P_e}[D_s D_t^\top] &= \mathbb{E}_{P_e}\left[D_s \mathbb{E}_{P_e}[D_t^\top \mid \mathcal{F}_t]\right] = 0, \\ \mathbb{E}_{P_e}[SS^\top] &= \sum_{t=0}^{H-1}\mathbb{E}_{P_e}[D_t D_t^\top] + \sum_{0\le s<t<H}\mathbb{E}_{P_e}[D_s D_t^\top + D_t D_s^\top] \\ &= \sum_{t=0}^{H-1}\mathbb{E}_{P_e}[D_t D_t^\top].\end{aligned} \tag{55}$$

Finally, for every deterministic unit vector u, orthogonality and Cauchy–Schwarz imply

$$\begin{aligned}u^\top \mathcal{I}_e u &= \sum_{t=0}^{H-1}\mathbb{E}_{P_e}[(u^\top D_t)^2] \\ &\le \sum_{t=0}^{H-1}\mathbb{E}_{P_e}[\|u\|_2^2\|D_t\|_2^2] \\ &\le G^2\sum_{t=0}^{H-1}P_e(t<T) = G^2\mathbb{E}_{P_e}T \le HG^2.\end{aligned} \tag{56}$$

Taking the supremum over unit vectors proves the eigenvalue bound. The almost-sure score bound follows separately from the triangle inequality, $\|S\|_2 \le \sum_{t<T}\|D_t\|_2 \le GT$. □

Lemma 5 (Two successive changes of measure)**.** *Let $P_0 = p_{\beta,v}$, $P = p_{\theta,v}$, and $Q = p_{\theta,v_\star}$. Assume the source policy covers the current policy on the relevant histories, and the environment model and overlap assumptions of Theorem 3 hold. Then*

$$\begin{aligned}\frac{dP}{dP_0}(\tau) &= A(\tau) = \prod_{t<T} \frac{\pi_\theta(a_t \mid h_t)}{\beta(a_t \mid h_t)},\\ \frac{dQ}{dP}(\tau) &= L(\tau) = \prod_{t<T} \rho_v^E(u_t, z_t),\\ \frac{dQ}{dP_0}(\tau) &= W(\tau) = A(\tau)L(\tau).\end{aligned} \tag{57}$$

Consequently $\mathbb{E}_{P_0}[WF] = \mathbb{E}_Q F$ for every Q-integrable vector-valued function F, and W is the unique nonnegative weight, up to P_0-null sets, that recovers all bounded measurable scalar trajectory functionals.

Proof. The source and target transition laws have disjoint, observed label supports. Hence on the support of label y, their common within-label kernel cancels, giving

$$\begin{aligned}P_{v_\star}(dz \mid u) &= \rho_v^E(u,z) P_v(dz \mid u),\\ \rho_v^E(u,z) &= \frac{p_{v_\star}(y(z) \mid k(u))}{p_v(y(z) \mid k(u))}, \qquad \int \rho_v^E(u,z) P_v(dz \mid u) = 1.\end{aligned} \tag{58}$$

For unaffected and padded transitions the ratio is one. Source probabilities are strictly positive on every label that can occur under the target law, so this kernel identity entails absolute continuity. Similarly, source-policy coverage gives the conditional identity $\pi_\theta(da \mid h) = [\pi_\theta(a \mid h)/\beta(a \mid h)]\beta(da \mid h)$. The common initial law contributes no factor.

The padded horizon and iterated integration extend the trajectory identity to general state spaces. Let $\tau_{0:t}$ denote the partial trajectory just before action t, and let f be any nonnegative measurable function of the next extended history. At each active step, the action and transition factors satisfy

$$\begin{aligned}&\int \beta(da \mid h_t) P_v(dz \mid u) \frac{\pi_\theta(a \mid h_t)}{\beta(a \mid h_t)} \rho_v^E(u,z) f(\tau_{0:t}, a, z)\\ &\quad = \int \pi_\theta(da \mid h_t) P_v(dz \mid u) \rho_v^E(u,z) f(\tau_{0:t}, a, z)\\ &\quad = \int \pi_\theta(da \mid h_t) P_{v_\star}(dz \mid u) f(\tau_{0:t}, a, z).\end{aligned} \tag{59}$$

Tonelli's theorem permits successive applications of this equality through the H steps. The resulting product identity establishes $Q \ll P_0$ and $dQ/dP_0 = AL$. Omitting the transition ratios in the same argument proves $dP/dP_0 = A$; using the current policy on both sides proves $dQ/dP = L$. This also proves the Radon–Nikodym chain in Equation (57), including zero target factors.

For a vector-valued F with $\mathbb{E}_Q\|F\|_2 < \infty$, apply the nonnegative scalar identity first to $\|F\|_2$ and then to the positive and negative parts of every coordinate. Thus

$$\begin{aligned}\mathbb{E}_{P_0}[W\|F\|_2] &= \mathbb{E}_Q\|F\|_2 < \infty,\\ \mathbb{E}_{P_0}[WF] &= \mathbb{E}_Q F.\end{aligned} \tag{60}$$

If another measurable $\bar{W} \geq 0$ recovers every bounded scalar functional, choose indicator functions of arbitrary measurable trajectory events B. Then

$$\int_B \bar{W}\, dP_0 = \mathbb{E}_{P_0}[\bar{W}\mathbf{1}_B] = Q(B) = \mathbb{E}_{P_0}[W\mathbf{1}_B] = \int_B W\, dP_0. \tag{61}$$

Taking B to be the sets where $\bar{W} - W$ is positive or negative proves $\bar{W} = W$ almost surely. Uniqueness applies to recovery of all bounded trajectory tests; an inexact weight may still recover an individual reward gradient. □

Proof of Theorem 3. Lemma 4 provides the target likelihood-ratio gradient identity, while Lemma 5 supplies the exact joint change of measure. Combining them yields

$$\begin{aligned}\mathbb{E}_{P_0}[W(R-b)S] &= \mathbb{E}_Q[(R-b)S] = \mathbb{E}_Q[RS] - b\mathbb{E}_Q S = \nabla J_\star(\theta) = g_\star,\\ \mathbb{E}_{P_0}[A(R-b)S] &= \mathbb{E}_P[(R-b)S] = \mathbb{E}_P[RS] - b\mathbb{E}_P S = \nabla J_v(\theta) = g_v.\end{aligned} \tag{62}$$

Thus policy-only replay computes the source-environment gradient even when it perfectly corrects the behavior policy. The product AL corrects both laws.

The detached replay loss yields the same weighted score. Let ϑ be the optimization variable while θ remains the frozen scoring parameter, and define

$$\begin{aligned}\ell(\vartheta;\tau) &= -\operatorname{sg}(W(\tau)(R(\tau)-b))\sum_{t<T}\log\pi_\vartheta(a_t \mid h_t),\\ -\nabla_\vartheta \ell(\vartheta;\tau)\big|_{\vartheta=\theta} &= W(\tau)(R(\tau)-b)S(\tau),\\ \mathbb{E}_{P_0}\big[-\nabla_\vartheta \ell(\vartheta;\tau)\big|_{\vartheta=\theta}\big] &= g_\star.\end{aligned} \tag{63}$$

The weight is held fixed during differentiation, keeping the surrogate gradient equal to the weighted score.

The identity also yields consistency of an independent Monte Carlo replay mean. For independent, identically distributed $\tau_i \sim P_0$, define $\widehat{g}_n = n^{-1}\sum_{i=1}^n W_i(R_i-b)S_i$. Its first absolute moment is finite because

$$\begin{aligned}\mathbb{E}_{P_0}\|W(R-b)S\|_2 &= \mathbb{E}_Q[|R-b|\|S\|_2]\\ &\le (1+|b|)\mathbb{E}_Q\|S\|_2\\ &\le (1+|b|)G\mathbb{E}_Q T \le (1+|b|)GH < \infty.\end{aligned} \tag{64}$$

The scalar strong law applied to each of the finitely many parameter coordinates therefore gives

$$\widehat{g}_n \longrightarrow g_\star \qquad \text{almost surely as } n\to\infty. \tag{65}$$

This consistency result requires only the first absolute moment of $W(R-b)S$, and uses exact, uncapped ratios on all source records. For a fixed source mixture with component probabilities η_j, each component must cover the target; conditioning on the source index and applying the preceding identity gives

$$\sum_j \eta_j \mathbb{E}_{p_{\beta_j,v_j}}[W_j(R-b)S] = \sum_j \eta_j g_\star = g_\star, \qquad \sum_j \eta_j = 1. \tag{66}$$

An independent fresh mean and an exact replay mean consequently retain expectation $g_\star$ when combined with the fixed coefficients $1/2$ and $1/2$. Conditioning on a realized FIFO buffer yields a weighted empirical mean. Applying the population identity requires independent fixed-source sampling; policy selection from the same buffer introduces dependence even with task folds.

It remains to bound the difference between the policy-only and corrected population gradients. Write $\Delta_g = g_\star - g_v$ and let $\mu = Q - P$ be the finite signed difference measure. Both score expectations vanish by Lemma 4; hence subtracting half the score changes neither gradient difference:

$$\begin{aligned}\Delta_g &= \int RS\,d\mu = \int (R-\tfrac{1}{2})S\,d\mu,\\ \int S\,d\mu &= \mathbb{E}_Q S - \mathbb{E}_P S = 0.\end{aligned} \tag{67}$$

Here total variation is defined as $\mathrm{TV}(P,Q) = \sup_B |P(B)-Q(B)|$, so the total mass of the variation measure $|\mu|$ is $2\,\mathrm{TV}(P,Q)$. If M is an essential upper bound for $\|S\|_2$ under both laws, norm duality and the variation inequality give

$$\begin{aligned}\|\Delta_g\|_2 &= \sup_{\|u\|_2=1}\left|\int (R-\tfrac{1}{2})u^\top S\,d\mu\right|\\ &\le \sup_{\|u\|_2=1}\int |R-\tfrac{1}{2}|\,|u^\top S|\,d|\mu|\\ &\le \frac{1}{2}\sup_{\|u\|_2=1}\int \|u\|_2\|S\|_2\,d|\mu|\\ &\le \frac{M}{2}\int d|\mu| = M\,\mathrm{TV}(P,Q).\end{aligned} \tag{68}$$

The factor $1/2$ in reward centering is essential: using RS directly would lose a factor of two. Proposition 1 gives an overlapping two-tool instance attaining Equation (68) with nonzero bias.

For the Fisher bound, suppose $\chi^2(Q\|P) = \mathbb{E}_P[(L-1)^2] < \infty$. Put $V = L - 1$. By the Radon–Nikodym identity, $\mathbb{E}_P V = 0$ and $\mathbb{E}_P[VS] = \mathbb{E}_Q S - \mathbb{E}_P S = 0$. For every deterministic unit vector u, Cauchy–Schwarz and the bounded reward give

$$\begin{aligned} |u^\top \Delta_g| &= \left|\mathbb{E}_P[V(R - \tfrac{1}{2})u^\top S]\right| \\ &\le \mathbb{E}_P[|V|\,|R - \tfrac{1}{2}|\,|u^\top S|] \\ &\le \left(\mathbb{E}_P V^2\right)^{1/2}\left(\mathbb{E}_P[(R-\tfrac{1}{2})^2(u^\top S)^2]\right)^{1/2} \\ &\le \frac{1}{2}\sqrt{\chi^2(Q\|P)}\left(\mathbb{E}_P[(u^\top S)^2]\right)^{1/2} \\ &= \frac{1}{2}\sqrt{\chi^2(Q\|P)}\sqrt{u^\top I_v u} \\ &\le \frac{1}{2}\sqrt{\chi^2(Q\|P)\lambda_{\max}(I_v)}. \end{aligned} \tag{69}$$

Taking the supremum over u and combining Equations (68) and (69) proves

$$\begin{aligned} \|g_\star - g_v\|_2 &\le \min\left\{ M\,\mathrm{TV}(P,Q),\ \frac{1}{2}\sqrt{\chi^2(Q\|P)\lambda_{\max}(I_v)} \right\} \\ &\le \min\left\{ HG\,\mathrm{TV}(P,Q),\ \frac{G}{2}\sqrt{\chi^2(Q\|P)\mathbb{E}_P T} \right\}. \end{aligned} \tag{70}$$

If the chi-square divergence is infinite, the total-variation part remains valid and the Fisher part is simply omitted. If $I_v = 0$, then $S = 0$ P-almost surely and also Q-almost surely, so both gradients are zero directly. The nontrivial Fisher constant $1/2$ is attained by the second construction in Proposition 1. Both constants are therefore optimal under the stated assumptions. The divergence can still grow with the horizon, making replay accuracy depend on trajectory length. □

Proof of Theorem 1. By the positive tool-anchor and source-label probabilities in Appendix D.1, every tool–outcome cell has positive source mass. Under Equation (2), the conditional density is constant on each such cell. Theorem 2 therefore identifies the unregularized population correction with ρ_v^E; its validity at every restored anchor of a covered tool follows from the state-independent mixing probabilities. Thus the product L in the main theorem is the exact environment likelihood ratio, including factors equal to one at unaffected or absorbing transitions. Theorem 3, with $S = \mathsf{s}_\theta$, now gives both identities in Equation (10) and almost-sure convergence of the corrected empirical mean. Source trajectories with $A = 0$ have zero weighted contribution by the stated convention, so no score value outside the current policy's support is needed. □

Corollary 1 (A sharper bound from exact score control variates)**.** *Under Theorem 3, assume $\chi^2(Q\|P) < \infty$. Let $Y = RS$, $m = \mathbb{E}_P Y$, $B = \mathbb{E}_P[(Y-m)S^\top] = \mathbb{E}_P[RSS^\top]$, and define*

$$\begin{aligned} Z &= Y - m - BI_v^\dagger S, \\ \mathcal{C}_{\mathrm{res}} &= \mathrm{Cov}_P(Y) - BI_v^\dagger B^\top. \end{aligned} \tag{71}$$

Then $\mathcal{C}_{\mathrm{res}} = \mathbb{E}_P[ZZ^\top] \succeq 0$ and

$$\begin{aligned} \|g_\star - g_v\|_2 &\le \min\left\{ M\,\mathrm{TV}(P,Q),\ \sqrt{\chi^2(Q\|P)\lambda_{\max}(\mathcal{C}_{\mathrm{res}})} \right\}, \\ \mathcal{C}_{\mathrm{res}} &\preceq \mathbb{E}_P[(R-\tfrac{1}{2})^2 SS^\top] \preceq \tfrac{1}{4} I_v. \end{aligned} \tag{72}$$

Moreover, Z has the smallest second-moment matrix, in the positive-semidefinite order, among residuals $Y - a - CS$ with deterministic vector a and matrix C. Thus the bound removes every exact control variate in the affine span of the trajectory score before applying Cauchy–Schwarz. These control variates sharpen the population analysis; ESR *uses the estimator in Section 4.2.*

Proof. The pseudoinverse is needed because the score covariance can be singular. If $u \in \ker I_v$, then $\mathbb{E}_P[(u^\top S)^2] = 0$, so $u^\top S = 0$ almost surely and $Bu = 0$. Consequently $B = BI_v^\dagger I_v$. This

establishes the orthogonality relations

$$\begin{aligned}\mathbb{E}_P Z &= \mathbb{E}_P Y - m - BI_v^\dagger \mathbb{E}_P S = 0,\\ \mathbb{E}_P[ZS^\top] &= B - BI_v^\dagger I_v = 0.\end{aligned} \tag{73}$$

Expanding the second moment, using the symmetry of $I_v^\dagger$, and using $I_v^\dagger I_v I_v^\dagger = I_v^\dagger$, gives

$$\begin{aligned}\mathbb{E}_P[ZZ^\top] &= \mathrm{Cov}_P(Y) - BI_v^\dagger B^\top - BI_v^\dagger B^\top + BI_v^\dagger I_v I_v^\dagger B^\top\\ &= \mathrm{Cov}_P(Y) - BI_v^\dagger B^\top = \mathcal{C}_{\mathrm{res}} \succeq 0.\end{aligned} \tag{74}$$

For arbitrary a and C, write $Y - a - CS = Z + (m-a) + (BI_v^\dagger - C)S$. All cross terms vanish by Equation (73) and $\mathbb{E}_P S = 0$. The resulting matrix Pythagorean identity is

$$\begin{aligned}&\mathbb{E}_P[(Y-a-CS)(Y-a-CS)^\top]\\ &\quad= \mathcal{C}_{\mathrm{res}} + (m-a)(m-a)^\top + (BI_v^\dagger - C)I_v(BI_v^\dagger - C)^\top \succeq \mathcal{C}_{\mathrm{res}}.\end{aligned} \tag{75}$$

This proves optimality for both singular and nonsingular covariances. Setting $a = 0$ and $C = \frac{1}{2}I$ in this identity and then testing the resulting matrices against an arbitrary vector u gives

$$\begin{aligned}u^\top \mathcal{C}_{\mathrm{res}} u &\le u^\top \mathbb{E}_P[(Y - \tfrac{1}{2}S)(Y-\tfrac{1}{2}S)^\top]u\\ &= \mathbb{E}_P[(R-\tfrac{1}{2})^2 (u^\top S)^2]\\ &\le \frac{1}{4}\mathbb{E}_P[(u^\top S)^2] = \frac{1}{4}u^\top I_v u.\end{aligned} \tag{76}$$

Since $V = L - 1$ is orthogonal to both constants and the score, replacing Y by Z does not change the gradient difference. A final directional Cauchy–Schwarz chain yields

$$\begin{aligned}|u^\top \Delta_g| &= \left|\mathbb{E}_P[Vu^\top Y]\right| = \left|\mathbb{E}_P[Vu^\top Z]\right|\\ &\le \sqrt{\mathbb{E}_P V^2}\sqrt{\mathbb{E}_P[(u^\top Z)^2]}\\ &= \sqrt{\chi^2(Q\|P)}\sqrt{u^\top \mathcal{C}_{\mathrm{res}} u}\\ &\le \sqrt{\chi^2(Q\|P)\lambda_{\max}(\mathcal{C}_{\mathrm{res}})} \qquad (\|u\|_2 = 1).\end{aligned} \tag{77}$$

Norm duality proves the refined bound. Equation (76) shows that it is at least as tight as the Fisher bound in Theorem 3. □

D.5 Policy-gradient identity and correction error

Fix the current parameters θ_n, a behavior policy β, a source version v, and a baseline b independently of the trajectories under consideration. Write $p_0 = p_{\beta,v}$ and $p_\star = p_{\theta_n,v_\star}$. Under $p_\star \ll p_0$, cancellation of the common initial law gives Equation (3), and $\mathbb{E}_{p_0}[WF] = \mathbb{E}_{p_\star}F$ for every integrable F. Differentiating the target trajectory law gives

$$\nabla J_\star(\theta_n) = \mathbb{E}_{p_0}\left[W(\tau)(R(\tau)-b)\sum_{t<T}\nabla_\theta \log \pi_\theta(a_t \mid h_t)\big|_{\theta=\theta_n}\right]. \tag{78}$$

The baseline term vanishes because the expected score of $p_\star$ is zero. The negative gradient of the detached-weight replay loss at θ_n has this expectation. Applying the identity to each component of a source mixture requires target coverage by each component. Two task folds separate a scored task from its probes. The identity still requires the policy to be fixed independently of the source trajectories; selecting it using the same finite buffer introduces dependence.

Evidence rejection, estimation, and capping. To include missing support, write the Lebesgue decomposition $p_\star = Wp_0 + p_\perp$, where $p_\perp$ has mass δ. Let $g(\tau) \in \{0,1\}$ indicate a valid trajectory whose changed transitions pass all evidence checks. Let $\widehat{W}$ be the positive estimated product on $g = 1$, assigning it any nonnegative value on $g = 0$. The implemented weight is $\widetilde{W} = g\min(c, \widehat{W})$. For a vector-valued F with $\|F(\tau)\|_2 \le M$,

$$\begin{aligned}\left\|\mathbb{E}_{p_0}[\widetilde{W}F] - \mathbb{E}_{p_\star}F\right\|_2 \le M\big(&\mathbb{E}_{p_0}[g|\widehat{W} - W|] + \mathbb{E}_{p_0}[g(\widehat{W}-c)_+]\\ &+ \mathbb{E}_{p_0}[(1-g)W] + \delta\big).\end{aligned} \tag{79}$$

Indeed, $\widetilde{W} - W = g(\widehat{W} - W) - g(\widehat{W} - c)_+ - (1-g)W$. Apply the triangle inequality and bound $\| \int F \, dp_\perp \|_2$ by $M\delta$. The four terms isolate estimation error, capping, excluded target mass on source support, and missing support. The evidence rule controls where estimates are used, at the cost of the third term. With an unbiased fresh component and a fixed half mixture, the population bias of the combined update is half the replay bias.

For the policy-gradient integrand, $R, b \in [0,1]$, $T \le H$, and a per-turn score bound $\|\nabla \log \pi_\theta(a_t \mid h_t)\|_2 \le G$ imply $M = HG$. Equation (79) applies to bounded integrands. Local errors can grow with trajectory length: writing $A = \prod_t \pi_{\theta_n}(a_t \mid h_t)/\beta(a_t \mid h_t)$, $e_t = \rho_t^E$, and $\widehat{e}_t = \widehat{\rho}_t^E$ on an eligible trajectory gives

$$\widehat{W} - W = A \sum_{k=0}^{T-1} (\widehat{e}_k - e_k) \left(\prod_{j<k} \widehat{e}_j \right) \left(\prod_{j>k} e_j \right). \tag{80}$$

This follows by replacing the factors one at a time. If $A \le A_{\max}$ and all environment factors are bounded by $L \ge 1$, then $|\widehat{W} - W| \le A_{\max} L^{H-1} \sum_k |\widehat{e}_k - e_k|$. Even exact weights can have large variance: with a matched policy and independent transition ratios of source mean one and variance ζ, the trajectory weight variance is $(1+\zeta)^T - 1$. Thus fitting is computationally linear while statistical error can grow with the horizon.

D.6 Quantitative consequences of uncorrected reliability drift

Corollary 2 (A drift certificate from tool reliability)**.** *Let $a_k = p_v(1 \mid k) \in (0,1)$, $b_k = p_{v_\star}(1 \mid k)$, and let $\varepsilon_t \in [0,1]$ bound $|b_k - a_k|$ for every affected tool that can be invoked at turn t. Unaffected and absorbing transitions contribute zero. With* $\mathrm{kl}(b\|a) = b\log(b/a) + (1-b)\log((1-b)/(1-a))$,

$$\begin{aligned} \|g_\star - g_v\|_2 &\le M \min\left\{ 1 - \prod_{t=0}^{H-1} (1-\varepsilon_t), \ \sqrt{\tfrac{1}{2}\mathbb{E}_Q \sum_{t<T:\,\text{affected}} \mathrm{kl}(b_{k_t}\|a_{k_t})} \right\}, \\ 1 - \prod_{t=0}^{H-1} (1-\varepsilon_t) &\le \min\left\{ 1, \sum_{t=0}^{H-1} \varepsilon_t \right\}. \end{aligned} \tag{81}$$

The product bound on trajectory total variation is attainable; it requires no independence between the agent's successive tool choices.

Throughout this subsection, $P = p_{\theta,v}$ and $Q = p_{\theta,v_\star}$ use the same fixed policy and the same initial law. We assume the measurable state, observation, and action spaces are standard Borel, so the conditional kernels and couplings used below exist. Write $a_k = p_v(1 \mid k) \in (0,1)$ and $b_k = p_{v_\star}(1 \mid k) \in [0,1]$. The version-independent conditional kernels have the disjoint, label-identified supports required by Equation (2). A deterministic absorbing state pads every stopped trajectory to H transitions; padded transitions have zero drift and zero relative entropy. Thus padding changes neither the trajectory law on the original records nor any reward or score. At an active affected call, $k_t = k(s_t, a_t)$; expressions indexed by k_t are defined as zero at unaffected or padded transitions. All statements below concern the complete trajectory, including the latent environment state.

Exact one-step discrepancies. Fix a pre-transition pair $u = (s,a)$ invoking tool k, and abbreviate $a = a_k$, $b = b_k$, $P_u = P_v(\cdot \mid u)$, and $Q_u = P_{v_\star}(\cdot \mid u)$. Let $B_y(u)$ be the measurable set with reliability label y. The support assumption gives $G_y(B_y(u) \mid u) = 1$ and $G_{1-y}(B_y(u) \mid u) = 0$. Since $a \in (0,1)$, $Q_u \ll P_u$, with

$$\frac{dQ_u}{dP_u}(z) = \begin{cases} (1-b)/(1-a), & z \in B_0(u), \\ b/a, & z \in B_1(u), \end{cases} \qquad P_u\text{-almost surely.} \tag{82}$$

Define total variation as $\mathrm{TV}(\mu,\nu) = \sup_A |\mu(A) - \nu(A)|$. For every measurable A, the common conditional kernels imply

$$\begin{aligned} |Q_u(A) - P_u(A)| &= |b-a| \, |G_1(A \mid u) - G_0(A \mid u)| \\ &\le |b-a|. \end{aligned} \tag{83}$$

Taking $A = B_1(u)$ attains the last bound. Consequently, with the conventions $0 \log 0 = 0$ and $0 \log(0/c) = 0$ for $c > 0$,

$$\begin{aligned} \mathrm{TV}(P_u, Q_u) &= |b - a|, \\ D_{\mathrm{KL}}(Q_u \| P_u) &= \int \log\left(\frac{dQ_u}{dP_u}\right) dQ_u \\ &= b \log \frac{b}{a} + (1-b) \log \frac{1-b}{1-a} = \mathrm{kl}(b \| a). \end{aligned} \tag{84}$$

Both discrepancies depend exactly on the label probabilities, with the shared within-label response laws canceling out.

Coupling complete trajectories under adaptive tool selection. For each $t \in \{0, \dots, H-1\}$, let $\varepsilon_t \in [0,1]$ bound $|b_{k_t} - a_{k_t}|$ for every affected tool that can be invoked at that step; unaffected and absorbing transitions have discrepancy zero. We construct a coupling (τ^-, τ^+) of P and Q sequentially. Initially, draw one common (x, s_0, o_0) from q. Whenever the two complete prefixes agree, their observable histories agree, so draw a single common action from $\pi_\theta(\cdot \mid h_t)$. For an affected call at the common pair u, draw U_t uniformly on $[0,1]$ and set

$$Y_t^- = \mathbf{1}\{U_t \le a_{k_t}\}, \qquad Y_t^+ = \mathbf{1}\{U_t \le b_{k_t}\}, \qquad \Pr(Y_t^- \ne Y_t^+ \mid u) = |b_{k_t} - a_{k_t}|. \tag{85}$$

If the labels agree at y, draw one common next state–observation pair from $G_y(\cdot \mid u)$ and assign it to both branches. If the labels differ, draw each branch from its corresponding G_y; the disjoint supports force the two next pairs to differ. For an unaffected or padded transition, use one common draw from the common transition kernel. After the first disagreement, complete the two branches using any coupling with the prescribed marginal kernels, for example conditional independence. This construction preserves the correct marginal law of each complete trajectory. It allows dependent tool outcomes across time and a state-dependent policy.

Let E_t be the event that the two complete prefixes agree through the first t transitions, and let $\mathcal{C}_t$ denote the coupled prefix together with the common action at step t whenever E_t holds. The preceding construction and the tower property give

$$\begin{aligned} \Pr(E_{t+1}) &= \mathbb{E}[\mathbf{1}_{E_t} \Pr(E_{t+1} \mid \mathcal{C}_t)] \\ &= \mathbb{E}[\mathbf{1}_{E_t} (1 - |b_{k_t} - a_{k_t}|)] \\ &\ge (1 - \varepsilon_t) \Pr(E_t) \\ &\ge (1 - \varepsilon_t) \prod_{j=0}^{t-1} (1 - \varepsilon_j) = \prod_{j=0}^{t} (1 - \varepsilon_j). \end{aligned} \tag{86}$$

Here the discrepancy in the second line is zero for unaffected or padded steps; the final inequality follows inductively from $\Pr(E_0) = 1$. The product follows from lower bounds on conditional survival probabilities, allowing dependence between steps.

For any measurable trajectory event A, the coupling yields

$$\begin{aligned} |Q(A) - P(A)| &= \left|\mathbb{E}\left[\mathbf{1}_A(\tau^+) - \mathbf{1}_A(\tau^-)\right]\right| \\ &\le \mathbb{E}\left[\left|\mathbf{1}_A(\tau^+) - \mathbf{1}_A(\tau^-)\right|\right] \\ &\le \Pr(\tau^+ \ne \tau^-) \\ &= 1 - \Pr(E_H) \\ &\le 1 - \prod_{t=0}^{H-1} (1 - \varepsilon_t). \end{aligned} \tag{87}$$

Taking the supremum over A proves the product bound. Expanding the product gives a simpler, generally looser bound:

$$\begin{aligned}\mathrm{TV}(P,Q) &\le 1-\prod_{t=0}^{H-1}(1-\varepsilon_t)\\ &=\sum_{t=0}^{H-1}\varepsilon_t\prod_{j=0}^{t-1}(1-\varepsilon_j)\\ &\le \sum_{t=0}^{H-1}\varepsilon_t \le H\max_{0\le t<H}\varepsilon_t.\end{aligned} \tag{88}$$

The empty product is one. The first bound is always at most one, even when the last two bounds exceed one.

The exact relative-entropy chain rule. The common initial law and policy factors cancel in the complete-trajectory likelihood ratio. Positive source probabilities for both labels ensure $Q \ll P$, including when a target label probability equals zero. Let L_t denote the conditional ratio in Equation (82) at an affected call, and set $L_t = 1$ otherwise. Then, Q-almost surely,

$$\frac{dQ}{dP}(\tau)=\prod_{t=0}^{H-1}L_t, \qquad \log\frac{dQ}{dP}(\tau)=\sum_{t=0}^{H-1}\log L_t. \tag{89}$$

Each Q-positive transition has a positive likelihood ratio. Since there are finitely many tools and $H<\infty$, the relevant positive label probabilities and likelihood ratios range over finite sets, making the expectations below finite. Let $\mathcal{F}_t$ contain the complete target prefix and the sampled action at step t. Repeated conditional expectation and Equation (84) give

$$\begin{aligned}D_{\mathrm{KL}}(Q\|P) &= \mathbb{E}_Q\left[\log\frac{dQ}{dP}(\tau)\right]\\ &=\sum_{t=0}^{H-1}\mathbb{E}_Q[\log L_t]\\ &=\sum_{t=0}^{H-1}\mathbb{E}_Q[\mathbb{E}_Q[\log L_t\mid\mathcal{F}_t]]\\ &=\sum_{t=0}^{H-1}\mathbb{E}_Q[D_{\mathrm{KL}}(P_{v_\star}(\cdot\mid s_t,a_t)\|P_v(\cdot\mid s_t,a_t))]\\ &=\mathbb{E}_Q\left[\sum_{t<T}\mathrm{kl}(b_{k_t}\|a_{k_t})\right].\end{aligned} \tag{90}$$

The expectation uses target occupancy because earlier reliability outcomes affect subsequent tool choices and stopping decisions.

The entropy-to-variation inequality. Consider probability measures $Q \ll P$ and write $L = dQ/dP$. If $D_{\mathrm{KL}}(Q\|P) = +\infty$, the desired inequality is immediate. Otherwise, set $A = \{L \ge 1\}$, $q_A = Q(A)$, and $p_A = P(A)$. Since $\int(L-1)dP = 0$, the positive and negative parts have equal integral. Hence

$$\begin{aligned}\mathrm{TV}(P,Q) &= \int_A (L-1)dP = q_A - p_A\\ &= \frac{1}{2}\int |L-1|dP.\end{aligned} \tag{91}$$

To verify the supremum characterization directly, for every event B one has $\int_B(L-1)dP \le \int_A(L-1)dP$ and $-\int_B(L-1)dP \le \int_{A^c}(1-L)dP$; both right-hand sides agree, and $B = A$ attains them.

If $q_A = p_A$, the claim is trivial. Otherwise, $0 < p_A < 1$ and $q_A > p_A$: $p_A = 0$ would imply $q_A = 0$ by absolute continuity, while $p_A = 1$ would force $q_A - p_A \le 0$. Apply Jensen's inequality separately

on A and A^c to the convex function $z \mapsto z \log z$, extended continuously at zero. The two-cell log-sum inequality gives

$$\begin{aligned} D_{\mathrm{KL}}(Q\|P) &= p_A \mathbb{E}_P[L \log L \mid A] + (1-p_A)\mathbb{E}_P[L \log L \mid A^c] \\ &\geq p_A \mathbb{E}_P[L \mid A] \log \mathbb{E}_P[L \mid A] \\ &\quad + (1-p_A)\mathbb{E}_P[L \mid A^c] \log \mathbb{E}_P[L \mid A^c] \\ &= q_A \log \frac{q_A}{p_A} + (1-q_A) \log \frac{1-q_A}{1-p_A} \\ &= \mathrm{kl}(q_A\|p_A). \end{aligned} \tag{92}$$

Fix $p \in (0,1)$ and put $f_p(q) = \mathrm{kl}(q\|p)$. Direct differentiation on $(0,1)$ gives

$$f_p(p) = f_p'(p) = 0, \qquad f_p''(q) = \frac{1}{q(1-q)} \geq 4, \tag{93}$$

where the inequality uses $q(1-q) \leq 1/4$. For $p \leq q < 1$, integrate the second derivative twice and interchange the two integrals over the triangular region to obtain

$$\begin{aligned} \mathrm{kl}(q\|p) &= \int_p^q \left(\int_p^u \frac{dr}{r(1-r)} \right) du \\ &= \int_p^q \frac{q-r}{r(1-r)}\, dr \\ &\geq 4 \int_p^q (q-r)\, dr \\ &= 2(q-p)^2. \end{aligned} \tag{94}$$

For $q = 1$, the same conclusion follows by continuity as q increases to one. Combining the event reduction, Jensen's inequality, and this curvature calculation proves

$$\begin{aligned} D_{\mathrm{KL}}(Q\|P) &\geq \mathrm{kl}(q_A\|p_A) \\ &\geq 2(q_A - p_A)^2 = 2\,\mathrm{TV}(P,Q)^2, \\ \mathrm{TV}(P,Q) &\leq \sqrt{\frac{1}{2} D_{\mathrm{KL}}(Q\|P)}. \end{aligned} \tag{95}$$

Combining the two trajectory bounds and the gradient consequence. Substitute the exact entropy identity into Equation (95) and combine it with the coupling bound. The result is

$$\mathrm{TV}(P,Q) \leq \min\left\{ 1 - \prod_{t=0}^{H-1} (1-\varepsilon_t),\ \sqrt{\frac{1}{2} \mathbb{E}_Q \left[\sum_{t<T} \mathrm{kl}(b_{k_t}\|a_{k_t}) \right]} \right\}. \tag{96}$$

For the policy-gradient bound, let $S_\theta(\tau) = \sum_{t<T} \nabla_\theta \log \pi_\theta(a_t \mid h_t)$ and suppose $\|S_\theta\|_2 \leq M$ almost surely under both laws. For $w \in \{v, v_\star\}$, write $J_w(\theta) = \mathbb{E}_{p_{\theta,w}} R$. Under the differentiation assumptions of Theorem 3, $\mathbb{E}_P S_\theta = \mathbb{E}_Q S_\theta = 0$. Therefore the source-only policy-gradient population update, with any fixed scalar baseline, equals $\nabla J_v(\theta)$, and recentering the common reward at $1/2$ gives the sharper chain

$$\begin{aligned} \|\nabla J_{v_\star}(\theta) - \nabla J_v(\theta)\|_2 &= \left\| \int (R - \tfrac{1}{2}) S_\theta\, d(Q-P) \right\|_2 \\ &\leq \int |R - \tfrac{1}{2}|\, \|S_\theta\|_2\, d|Q-P| \\ &\leq \frac{M}{2} \int d|Q-P| \\ &= M\,\mathrm{TV}(P,Q) \\ &\leq M \min\left\{ 1 - \prod_{t=0}^{H-1} (1-\varepsilon_t),\ \sqrt{\frac{1}{2} \mathbb{E}_Q \left[\sum_{t<T} \mathrm{kl}(b_{k_t}\|a_{k_t}) \right]} \right\}. \end{aligned} \tag{97}$$

Here $|Q - P|$ denotes the total-variation measure of the signed measure $Q - P$. The factor improves from the generic bounded-integrand value $2M$ to M by using $R \in [0, 1]$ and the zero-score identities under both environment laws. In particular, a per-turn score bound G permits $M = HG$.

Equality cases and optimal constants. The following construction attains the product bound exactly. Fix $0 < \varepsilon_t < 1$ and consider a deterministic schedule of H tool calls, using one tool per step if the values differ. Let its source success probabilities be $a_{k_t} = 1 - \varepsilon_t$, its target success probabilities be $b_{k_t} = 1$, and let the common within-label kernel append the label to the state. This equality construction uses independent source labels; the theorem also covers dependent tool choices. The target assigns probability one to the all-success trajectory τ_{all}, while

$$\begin{aligned} P(\{\tau_{\mathrm{all}}\}) &= \prod_{t=0}^{H-1} (1 - \varepsilon_t), \\ \mathrm{TV}(P, Q) &\geq Q(\{\tau_{\mathrm{all}}\}) - P(\{\tau_{\mathrm{all}}\}) \\ &= 1 - \prod_{t=0}^{H-1} (1 - \varepsilon_t) \\ &\geq \mathrm{TV}(P, Q). \end{aligned} \tag{98}$$

The final line uses the already proved upper bound, so every inequality in this display is an equality. Zero-drift steps can be inserted with the same source and target kernel; boundary values with $\varepsilon_t = 1$ are approached as limits while retaining strictly positive source probabilities. Thus the product bound is sharp given only the stepwise total-variation limits.

The constant $1/2$ in the entropy square root is also optimal as a universal constant. For a single binary tool with $a = 1/2$ and $b = 1/2 + h$, where $h \to 0$, direct expansion gives

$$\mathrm{TV}(P, Q) = |h|, \qquad D_{\mathrm{KL}}(Q\|P) = 2h^2 + O(h^4), \qquad \frac{\mathrm{TV}(P, Q)^2}{D_{\mathrm{KL}}(Q\|P)} \longrightarrow \frac{1}{2}. \tag{99}$$

The two constructions establish sharpness of the product bound and the universal entropy constant separately, each for its stated drift profile. Both bounds hold for state-dependent tool selection and stopping, because the coupling conditions on the complete prefix and the entropy calculation averages over the actual target occupancy. This completes the proof of Corollary 2.

D.7 Necessity, opposite gradients, and equality cases

Proposition 1 (On-policy data can prescribe the opposite gradient)**.** *There is a one-turn, two-tool problem with common reward $R = y$, full source support, and $\beta = \pi_\theta$ for which $g_\star = -g_v \neq 0$. Specifically, let $\pi_\theta(1) = \sigma(\theta)$ and let the source success probabilities be $(1/2 + \epsilon, 1/2 - \epsilon)$, with target probabilities swapped, where $0 < \epsilon < 1/2$. Then*

$$g_v = 2\epsilon\sigma(\theta)(1 - \sigma(\theta)), \qquad g_\star = -2\epsilon\sigma(\theta)(1 - \sigma(\theta)). \tag{100}$$

At $\theta = 0$, equality holds in the total-variation part of Equation (48)*. A second two-tool construction attains its Fisher–χ^2 constant.*

Proof of Proposition 1 and sharpness in Theorem 3. Consider one initial state, two actions identifying two tools, and a terminal binary observation Y. Both environment versions use the same conditional kernel given (a, Y): the next state is simply the terminal state labeled by (a, Y). The reward is $R = Y$. Thus the within-label kernels agree exactly, the label supports are disjoint, and the construction satisfies Equation (2). Let $s = \sigma(\theta)$, so $\pi_\theta(1) = s$ and $\pi_\theta(2) = 1 - s$. The two scalar action scores and their common bound are

$$S(1) = 1 - s, \qquad S(2) = -s, \qquad \max_a |S(a)| = \max\{s, 1 - s\} \leq 1. \tag{101}$$

All policy probabilities are strictly positive at finite θ and differentiable in a neighborhood. Set the behavior policy equal to the frozen current policy, so that $A = 1$ exactly.

For the first construction, choose $0 < \epsilon < 1/2$ and source success probabilities $a_1 = 1/2 + \epsilon$, $a_2 = 1/2 - \epsilon$. Swap these probabilities in the target: $b_1 = a_2$, $b_2 = a_1$. The objectives and gradients

are

$$\begin{aligned} J_v(\theta) &= s(1/2+\epsilon) + (1-s)(1/2-\epsilon) = 1/2 - \epsilon + 2\epsilon s, \\ J_\star(\theta) &= s(1/2-\epsilon) + (1-s)(1/2+\epsilon) = 1/2 + \epsilon - 2\epsilon s, \\ g_v &= 2\epsilon s(1-s) > 0, \qquad g_\star = -2\epsilon s(1-s) < 0, \\ |g_\star - g_v| &= 4\epsilon s(1-s). \end{aligned} \tag{102}$$

Thus the two gradients are exact negatives for every finite parameter. The event that an action is sampled is part of the trajectory, so its total-variation calculation must retain the action probabilities. For any $s \in (0,1)$ it gives

$$\begin{aligned} \mathrm{TV}(P,Q) &= \frac{1}{2}\sum_{a=1}^{2}\pi_\theta(a)\sum_{y=0}^{1}|p_v(y \mid a) - p_{v_\star}(y \mid a)| \\ &= \sum_{a=1}^{2}\pi_\theta(a)|a_a - b_a| = 2\epsilon. \end{aligned} \tag{103}$$

At $\theta = 0$, $s = 1/2$, the exact score envelope is $M = 1/2$ and $|g_\star - g_v| = \epsilon = M\,\mathrm{TV}(P,Q)$. The discrepancy arises solely from changed reliability: every source trajectory has positive probability and the policy is fixed. The complete environment ratio is $L(a,y) = p_{v_\star}(y \mid a)/p_v(y \mid a)$. Direct cancellation also checks the correction:

$$\begin{aligned} \mathbb{E}_P[L(R-b)S] &= \sum_{a,y}\pi_\theta(a)p_v(y \mid a)\frac{p_{v_\star}(y \mid a)}{p_v(y \mid a)}(y-b)S(a) \\ &= \sum_{a}\pi_\theta(a)\big(b_a - b\big)S(a) \\ &= s(1-s)(b_1 - b_2) = -2\epsilon s(1-s) = g_\star. \end{aligned} \tag{104}$$

For sharpness of the Fisher constant, keep $\theta = 0$ and take the source success probabilities to be $(1/2, 1/2)$, with target probabilities $(1/2+\epsilon, 1/2-\epsilon)$. The four source trajectories (a,y) have equal mass $1/4$. Their score values are $\pm 1/2$, and their likelihood ratios are $1+2\epsilon$ or $1-2\epsilon$. Hence every inequality in the directional Cauchy–Schwarz argument is attained:

$$\begin{aligned} g_v = 0, \qquad g_\star = \epsilon/2, \qquad I_v = \mathbb{E}_P S^2 = 1/4, \\ \chi^2(Q\|P) = \tfrac{1}{4}\sum_{a,y}(L(a,y)-1)^2 = 4\epsilon^2, \\ |g_\star - g_v| = \epsilon/2 = \tfrac{1}{2}\sqrt{(4\epsilon^2)(1/4)} = \tfrac{1}{2}\sqrt{\chi^2(Q\|P)\lambda_{\max}(I_v)}. \end{aligned} \tag{105}$$

For every $0 < \epsilon < 1/2$, these finite, strictly supported examples attain the two constants in Equation (48) separately. Thus uniform gradient recovery over the stated problem class requires environment correction, even though individual reward gradients may agree under particular changes. □

D.8 Information lost by the fixed features

Proposition 2 (Indistinguishable feature probes with different target rewards)**.** *For any $s \in (0,1)$, there are a source law and two dominated target laws that produce the same distribution of paired tool–outcome features for every sample size, while their expectations of a common $[0,1]$-valued reward differ by s. Every estimator $\widehat{t}$ using only these paired features, known source information, and independent randomization satisfies*

$$\max\left\{\mathbb{E}_1|\widehat{t} - \mathbb{E}_{Q_1}f|, \mathbb{E}_2|\widehat{t} - \mathbb{E}_{Q_2}f|\right\} \geq \frac{s}{2}. \tag{106}$$

The lower bound is attained. These targets violate only the invariance of the within-label response law, and therefore lie outside Equation (2)*.*

Proof. Use one fixed anchor and one tool. Its response alphabet consists of $\perp, a, b$, where $\perp$ is labeled failure and a, b are labeled success. Choose the source and two target distributions as follows:

$$\begin{aligned} Q^-(\perp,a,b) &= (1-s, s/2, s/2), \\ Q_1(\perp,a,b) &= (1-s, s, 0), \qquad Q_2(\perp,a,b) = (1-s, 0, s), \\ f(\perp) &= 0, \qquad f(a) = 1, \qquad f(b) = 0. \end{aligned} \tag{107}$$

Every response in either target has source support. Under each of Q^-, Q_1, Q_2, the observed feature is the failure coordinate with probability $1-s$ and the success coordinate with probability s. With independent probe-branch randomness, a single paired feature has the same product-Bernoulli law under both alternatives; hence so does any sequence of such pairs. The distribution of $\widehat{t}$ is consequently identical under the two alternatives, including any independent randomization. Denote this common law by $\mathcal{D}$. Since $\mathbb{E}_{Q_1} f = s$ and $\mathbb{E}_{Q_2} f = 0$, the triangle inequality gives

$$\begin{aligned} s &= |s-0| \\ &\le |\widehat{t}-s| + |\widehat{t}|, \\ s &\le \mathbb{E}_{\mathcal{D}}|\widehat{t}-s| + \mathbb{E}_{\mathcal{D}}|\widehat{t}| \\ &= \mathbb{E}_1|\widehat{t}-\mathbb{E}_{Q_1} f| + \mathbb{E}_2|\widehat{t}-\mathbb{E}_{Q_2} f| \\ &\le 2\max\left\{\mathbb{E}_1|\widehat{t}-\mathbb{E}_{Q_1} f|, \mathbb{E}_2|\widehat{t}-\mathbb{E}_{Q_2} f|\right\}. \end{aligned} \tag{108}$$

The constant estimator $\widehat{t} = s/2$ has loss $s/2$ in both alternatives, proving attainability even when s is known. The population ESR feature ratio is exactly one, since both source and target feature frequencies agree. Retaining whether the success response was a or b would distinguish the targets. The ambiguity arises from the fixed feature representation. □

Selection bias from outcome-dependent filtering. Outcome-dependent filtering can bias paired probes even with shared anchors. To see this, let independent source and target success indicators have probabilities $1/4$ and $3/4$, respectively, and retain a pair only when its two indicators agree. Conditional on retention, their two marginals become identical:

$$\begin{aligned} \Pr(Y^- = Y^+ = 1) &= (1/4)(3/4) = 3/16, \\ \Pr(Y^- = Y^+ = 0) &= (3/4)(1/4) = 3/16, \\ \Pr(Y^- = 1 \mid Y^- = Y^+) = \Pr(Y^+ = 1 \mid Y^- = Y^+) &= \frac{3/16}{6/16} = 1/2. \end{aligned} \tag{109}$$

Thus accepted frequencies would give the unit correction even though the true success ratio is 3 and the failure ratio is $1/3$. The retained-pair assumption in Appendix D.1 rules out this selection bias. Selecting accessible anchors before execution preserves the assumption when all valid wrapper outcomes are retained.

D.9 Missing support and the exact reweighting limit

Missing source support creates an irreducible worst-case reweighting error, even with exact knowledge of both laws. The theorem quantifies this error over all bounded functionals.

Theorem 4 (Exact limit of source-only reweighting)**.** *For arbitrary source and target trajectory laws, write $Q = hP_0 + Q_\perp$ with $Q_\perp \perp P_0$ and $\delta = Q_\perp(\Omega)$. For any integrable signed source weight w and $M_F > 0$, with $\|F\|_\infty$ taken under $P_0 + Q$,*

$$\begin{aligned} \sup_{\|F\|_\infty \le M_F} \|\mathbb{E}_Q F - \mathbb{E}_{P_0}[wF]\|_2 &= M_F\big(\mathbb{E}_{P_0}|h-w| + \delta\big), \\ \inf_{w \in L^1(P_0)} \sup_{\|F\|_\infty \le M_F} \|\mathbb{E}_Q F - \mathbb{E}_{P_0}[wF]\|_2 &= M_F\delta. \end{aligned} \tag{110}$$

If $w \ge 0$ is additionally constrained by $\mathbb{E}_{P_0} w = 1$, the infimum is $2M_F\delta$. All three constants are attained.

Proof of Theorem 4. The Lebesgue decomposition gives a nonnegative integrable h and a finite nonnegative measure $Q_\perp$ satisfying $\mathbb{E}_{P_0} h = 1 - \delta$. Mutual singularity yields a measurable set N with $P_0(N) = 0$ and $Q_\perp(N^c) = 0$. Fix any integrable signed weight w. For any measurable $\mathbb{R}^d$-valued F with pointwise norm at most M_F, where $d \ge 1$ is fixed, the difference of its expectations decomposes as

$$\mathbb{E}_Q F - \mathbb{E}_{P_0}[wF] = \int_{N^c} (h-w)F\, dP_0 + \int_N F\, dQ_\perp. \tag{111}$$

All integrals exist because h, w are integrable and F is bounded. The triangle inequality gives the upper bound:

$$\begin{aligned}\|\mathbb{E}_Q F - \mathbb{E}_{P_0}[wF]\|_2 &\le \left\| \int_{N^c} (h-w)F \, dP_0 \right\|_2 + \left\| \int_N F \, dQ_\perp \right\|_2 \\ &\le \int_{N^c} |h-w| \|F\|_2 \, dP_0 + \int_N \|F\|_2 \, dQ_\perp \\ &\le M_F \int |h-w| \, dP_0 + M_F Q_\perp(N) \\ &= M_F \big(\mathbb{E}_{P_0}|h-w| + \delta \big).\end{aligned} \tag{112}$$

To attain the bound, fix a unit vector $e \in \mathbb{R}^d$ and choose a single measurable test aligned with both parts of the signed error measure:

$$F_w(\tau) = \begin{cases} M_F \operatorname{sign}(h(\tau) - w(\tau)) e, & \tau \in N^c, \\ M_F e, & \tau \in N. \end{cases} \tag{113}$$

Take $\operatorname{sign}(0) = 0$. Substituting this test gives

$$\begin{aligned}\mathbb{E}_Q F_w - \mathbb{E}_{P_0}[wF_w] &= M_F \left(\int_{N^c} (h-w) \operatorname{sign}(h-w) \, dP_0 + Q_\perp(N) \right) e \\ &= M_F \big(\mathbb{E}_{P_0}|h-w| + \delta \big) e, \\ \|\mathbb{E}_Q F_w - \mathbb{E}_{P_0}[wF_w]\|_2 &= M_F \big(\mathbb{E}_{P_0}|h-w| + \delta \big).\end{aligned} \tag{114}$$

This proves the exact supremum for every w. Nonnegativity of $\mathbb{E}|h-w|$ then gives a lower bound $M_F\delta$ on the unrestricted infimum, and $w = h$ attains it. The same optimum therefore applies to signed and nonnegative weights.

For normalized nonnegative weights, the additional error is unavoidable:

$$\begin{aligned}\mathbb{E}_{P_0}|h-w| &\ge |\mathbb{E}_{P_0}(h-w)| \\ &= |(1-\delta) - 1| = \delta, \\ \sup_{\|F\|_\infty \le M_F} \|\mathbb{E}_Q F - \mathbb{E}_{P_0}[wF]\|_2 &= M_F \big(\mathbb{E}_{P_0}|h-w| + \delta \big) \ge 2M_F\delta.\end{aligned} \tag{115}$$

The choice $w = h + \delta$ is nonnegative and obeys $\mathbb{E}_{P_0} w = (1-\delta) + \delta = 1$, while $\mathbb{E}_{P_0}|h-w| = \delta$. It attains the lower bound even at the endpoints $\delta = 0$ or $\delta = 1$. The normalized statement concerns the stated bounded-test class; for scalar tests restricted to $[0,1]$, the corresponding distance between probability laws is half their signed-measure total-variation norm.

This worst-case bound concerns source-only reweighting: even an oracle w with complete knowledge of both laws leaves wP_0 with zero mass on N. Fresh trajectories, direct target evaluations, or a model can supply information on N. Individual score–reward integrands may be insensitive to this set; the displayed test class determines the exact worst-case error. □