DO SMALL ERRORS SNOWBALL? UNDERSTANDING AND CONTROLLING ERROR PROPAGATION IN LANGUAGE AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Small observation changes can propagate through a language agent's feedback loop, changing its actions, environment state, and later inputs. Terminal success conflates initial sensitivity with subsequent amplification, while one-step agreement cannot measure recovery. We introduce an affine Wasserstein profile in a policy-independent task-semantic geometry that separates leakage on equivalent states from amplification on separated states, with composition and first-exit bounds linking local dynamics to outcomes. Coupled-Rollout Contraction Optimization (CRCO) then trains marginal-preserving coupled continuations, penalizing successor separation only beyond a state-conditioned allowance. We prove that this correspondence distinguishes pair laws that absolute and constant-threshold penalties cannot, and that the optimized hinge bounds task-score sensitivity with a finite-sample guarantee. Across nine backbones and eight interactive settings, CRCO has the highest perturbed-score point estimate among six matched objectives on all nine backbones. It exceeds matched absolute propagation by 2.6/3.3 points on two instrumented anchors, retaining a 2.6-point mean gain with an observation-trained metric without latent predicates. Separately analyzed reference-prefix and reset-to-terminal protocols yield equal-cell mean gains of 3.2 and 2.5 points. Threshold reassignment, recovery interventions, and five-step checkpoint selection corroborate the mechanism; deployment remains one ordinary rollout.

1 INTRODUCTION

Language agents act through feedback loops: a small observation variation can change an action, the resulting environment state, and every subsequent input. Later interaction may repair or amplify that deviation, so long-horizon reliability is a joint property of the initial response and of recovery under feedback. Analyses of tool-using agents document such error propagation (Zhu et al., 2025; Gurram, 2026), and interactive benchmarks expose the persistent state and partial observability that determine whether recovery remains possible (Liu et al., 2023; Ma et al., 2024). Reliable long-horizon agents therefore require measuring how an injected deviation evolves and training policies whose feedback contracts it.

Two challenges prevent existing methods from controlling this propagation. **Challenge 1 (measurement):** existing robustness criteria conflate semantic sensitivity with feedback amplification. Terminal success mixes both effects, whereas one-step agreement captures only the initial injection; moreover, semantically equivalent states may induce different policy responses, creating representative-dependent leakage (Givan et al., 2003; Ferns et al., 2011). **Challenge 2 (optimization):** existing training penalties cannot distinguish recovery from expansion. An absolute successor-distance penalty assigns the same cost to a successor distance whether it follows a large or a nearly zero current separation, and independent branch sampling further mixes policy disagreement with sampling noise.

We make three contributions. **(1) Measurement:** an affine Wasserstein profile separates leakage on semantically equivalent states from amplification on separated states, and connects both to outcomes through perturbation, composition, and first-exit bounds (Sections 2–3). **(2) Optimization:** Coupled-Rollout Contraction Optimization (CRCO) trains maximally coupled continuations with a state-conditioned successor allowance (Section 4). Its hinge distinguishes pair laws with identical

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.

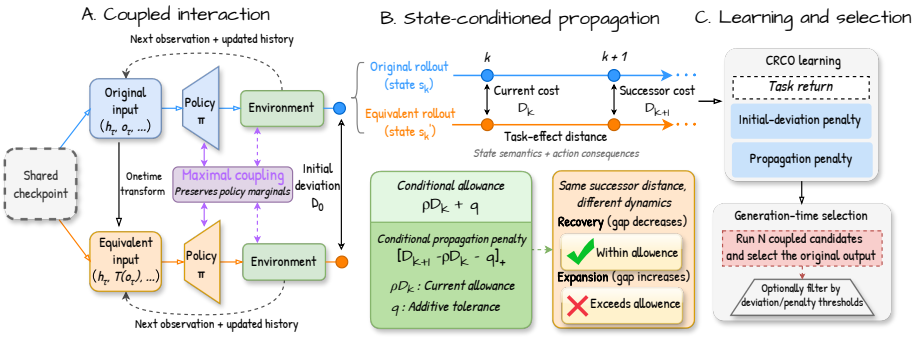


Figure 1: Overview. **A:** Two executions share a checkpoint and receive an observation or its admissible, semantically equivalent transform; maximal coupling preserves each policy marginal, and D_0 measures injection. **B:** Each successor distance D_{k+1} is compared with an allowance conditioned on the current distance D_k ($\bar{\rho}D_k + \bar{q}$ in Equation 16), so the same successor distance can constitute recovery or expansion. **C:** CRCO combines task return with initial-deviation and state-conditioned propagation penalties; the reported results deploy one ordinary rollout without generation-time selection.

distance marginals and bounds task-score sensitivity (Proposition 1; Theorem 2). **(3) Evaluation:** matched controls across nine backbones and eight settings test transfer, state correspondence, recovery, and checkpoint selection under two execution protocols (Section 5).

In summary, CRCO gains 2.6/3.3 perturbed-score points over injection- and compute-matched absolute propagation on Qwen2.5-7B/Gemma-3-12B, while Qwen's held-pair direct-state OT gain falls from 0.86 to 0.71. Reassigning current–successor thresholds across otherwise matched pairs lowers P by 2.8 points on ScienceWorld and 1.4 on WebShop, isolating the contribution of state correspondence. Analyzed as separate estimands, the reference-prefix block yields positive mean gains on all nine backbones and the reset-to-terminal block on eight, with equal-cell means of $+3.2$ and $+2.5$ points. On 28 held-out panels, five-step profiles select the test-oracle checkpoint 23 times (21 for full-horizon validation) with 40.7% fewer robustness calls, and deployment uses one ordinary rollout.

2 Semantic Closed-Loop Stability

We fix task semantics independently of the policy to distinguish injected deviations from subsequent amplification. Appendix B collects the notation.

2.1 Task-Semantic Geometry and Closed-Loop Dynamics

Let an interactive task be a partially observed controlled process $\mathcal{M} = (\mathcal{X}, \mathcal{A}, P, \mathcal{O}, Q, r, g)$ with latent state x_t, action a_t, transition P, renderer Q, progress reward r, and terminal success $g(x_H) \in [0, 1]$. The augmented pre-decision state is $z_t = (x_t, h_t, o_t) \in \mathcal{Z}$, with retained transcript/controller state h_t. We abbreviate $\Pi_\pi(\cdot \mid z_t) := \Pi_\pi(\cdot \mid h_t, o_t)$: the latent state x_t renders the analysis Markovian but never enters the model input. The post-decision state is $y_t = (z_t, a_t) \in \mathcal{Y} = \mathcal{Z} \times \mathcal{A}$.

The environment/history kernel $G(dz^+ \mid y)$ executes the current action, renders the next observation, and updates the retained history; the policy lifting kernel $S_\pi(dy^+ \mid z^+) = \delta_{z^+}(dz')\Pi_\pi(da^+ \mid z^+)$ then selects the next action. The post-decision Markov kernel is therefore the typed composition

$$K_\pi(dy^+ \mid y) = (GS_\pi)(dy^+ \mid y) = \int_{\mathcal{Z}} G(dz^+ \mid y) S_\pi(dy^+ \mid z^+). \tag{1}$$

Frozen task-semantic extractors $\phi_j : \mathcal{Z} \to \mathcal{E}_j$ cover latent-state, belief, retained-commitment, and affordance coordinates. With coordinate pseudometrics $\delta_j \in [0, 1]$ and training-frozen benchmark

weights $w_{bj} \geq 0$ summing to one,

$$d_Z(z, z') = \sum_{j \in \mathcal{J}} w_{bj}\, \delta_j(\phi_j(z), \phi_j(z'))\,. \tag{2}$$

The latent-state metric d_X is the renormalized restriction of d_Z to latent-state coordinates and never replaces d_Z in real-task profiles. Raw strings enter only through ϕ_j, so differently worded histories with identical task state, belief summary, commitments, and affordances are at distance zero. Appendix F specifies all coordinates and weights. For $y = (z, a)$ and fixed $\alpha \in [0, 1]$, define

$$d_Y(y, y') = \alpha d_Z(z, z') + (1 - \alpha)\mathcal{W}_{d_Z}(G(\cdot \mid z, a), G(\cdot \mid z', a'))\,. \tag{3}$$

Because the Wasserstein lift of a pseudometric is itself a pseudometric, d_Y is a pseudometric. It assigns zero cost to distinct action strings with identical task effects, whereas its transport term registers an action-induced deviation before that deviation is realized in the next state.

Along a paired rollout, write $Y_k = (Z_k, A_k)$ and let the coupled execution of (A_k, A'_k) produce (Z_{k+1}, Z'_{k+1}). We use the single-index convention

$$\widetilde{D}_k = \alpha d_Z(Z_k, Z'_k) + (1 - \alpha) d_Z(Z_{k+1}, Z'_{k+1}). \tag{4}$$

Hence $\widetilde{D}_k$ is observed after the step-k actions are executed. Conditional on (Y_k, Y'_k), its mean upper-bounds $d_Y(Y_k, Y'_k)$; the next sample $\widetilde{D}_{k+1}$ estimates the closed-loop successor cost defined below. For an admissible initial intervention, $\Delta_0 = d_Z(Z_0, Z'_0) = 0$, so $\widetilde{D}_0 = (1 - \alpha)\Delta_1$ and nonnegativity gives $\widetilde{D}_1 \geq \alpha \widetilde{D}_0/(1 - \alpha)$. We use $\alpha = 0.30$, selected from $\{0.20, 0.30, 0.40\}$ on development tasks. The resulting algebraic floor, $0.429\widetilde{D}_0$, lies below the training target $\bar{\rho} = 0.65$ and places 70% of the initial post-decision cost on the realized action effect.

2.2 Admissible interventions and error injection

An observation transformation $T \in \mathcal{T}$ is *admissible at z* if it preserves every task-relevant proposition and belief represented by the retained input, the set and semantics of valid actions, and the latent environment state, and if it leaves the retained history identical after tokenization and truncation. Examples include permutation of unordered JSON keys, reordering independent result fields, schema-preserving tool paraphrases, and fields generated independently of task-relevant hidden variables. Delayed, duplicated, corrupted, or belief-revealing feedback is inadmissible and evaluated separately.

At intervention time τ, two executions share (x_τ, h_τ); one receives o_τ and the other $T(o_\tau)$. Writing z_τ^T for the transformed pre-decision state, the induced post-decision laws are $\mu_0^T = S_\pi(\cdot \mid z_\tau)$ and $\nu_0^T = S_\pi(\cdot \mid z_\tau^T)$. The injected effect is

$$\epsilon_\tau(T) = \mathcal{W}_{d_Y}(\mu_0^T, \nu_0^T). \tag{5}$$

Injection therefore measures a change in the distribution of action consequences rather than in action strings. Because task semantics determine input equivalence independently of whether the policy is invariant, one geometry measures both the initial sensitivity and the later recovery.

2.3 Leakage and conditional amplification

The affine profile resolves both components of Challenge 1 in one inequality: zero-distance pairs identify semantic leakage, and positive-distance pairs identify residual amplification at a common leakage budget.

Definition 1 (Wasserstein quotient leakage and residual amplification)**.** *For a reachable admissible set $\mathcal{Y}_{\mathrm{adm}}$, define*

$$q_\pi^0 = \sup_{y, y' : d_Y(y, y') = 0} \mathcal{W}_{d_Y}(K_\pi(\cdot \mid y), K_\pi(\cdot \mid y'))\,. \tag{6}$$

For any independently fixed $q \geq q_\pi^0$, define

$$\rho_\pi(q) = \sup_{y, y' : d_Y(y, y') > 0} \frac{[\mathcal{W}_{d_Y}(K_\pi(\cdot \mid y), K_\pi(\cdot \mid y')) - q]_+}{d_Y(y, y')}. \tag{7}$$

The pair $(q, \rho_\pi(q))$ is a certifying Wasserstein profile. Comparisons use a common leakage budget or the nondominated leakage–amplification frontier.

For subsequent theory, fix one profile and abbreviate $(q_\pi, \rho_\pi) = (q, \rho_\pi(q))$. Every such profile satisfies

$$\mathcal{W}_{d_Y}(K_\pi(\cdot \mid y), K_\pi(\cdot \mid y')) \le q_\pi + \rho_\pi d_Y(y, y'). \tag{8}$$

At a fixed leakage budget, the loop is affinely contractive when $\rho_\pi < 1$, with radius $q_\pi/(1-\rho_\pi)$.

Executable coupling costs. Write $\mathcal{W}(y, y') = \mathcal{W}_{d_Y}(K_\pi(y), K_\pi(y'))$ and $\mathcal{C}_\gamma(y, y') = \mathbb{E}[\widetilde{D}_{k+1} \mid Y_k = y, Y'_k = y']$. For an executable successor coupling γ_K and environment coupling ξ_G, feasibility gives

$$\mathcal{W}(y, y') \le \mathbb{E}_{\gamma_K} d_Y(U, U') \le \mathcal{C}_\gamma(y, y'). \tag{9}$$

A coupled rollout therefore samples the executable upper bound, whereas empirical OT estimates the Wasserstein response itself; the two coincide only under optimal couplings at both layers. Appendix D.9 gives the full construction.

Empirical diagnostics. Hats denote fitted summaries under a specified probe law. We report the adjacent-cost slope $\widehat{b}_1$, zero-class leakage quantiles, and the direct-state conditional-OT gain $\widehat{g}_A$ on common input pairs. These are empirical diagnostics defined alongside the experiments (Section 5) and are not identified with the uniform theoretical profile (q_π, ρ_π).

3 From Local Stability to Trajectory Outcomes

Assumption 1 (Reachable Wasserstein affine contraction)**.** *On a forward-invariant reachable admissible set and at a leakage budget fixed independently of the positive-distance slope, K_π satisfies Equation 8 for finite $q_\pi, \rho_\pi \ge 0$.*

Theorem 1 (One-time perturbation)**.** *Suppose a one-time intervention produces post-decision laws μ_0, ν_0 with $\mathcal{W}_{d_Y}(\mu_0, \nu_0) \le \epsilon$. Under Assumption 1, for $k \ge 1$,*

$$\mathcal{W}_{d_Y}(\mu_0 K_\pi^k, \nu_0 K_\pi^k) \le \rho_\pi^k \epsilon + q_\pi \sum_{j=0}^{k-1} \rho_\pi^j. \tag{10}$$

For $\rho_\pi < 1$, the right side is $\rho_\pi^k \epsilon + q_\pi(1-\rho_\pi^k)/(1-\rho_\pi)$.

The proof appears in Appendix D.1. The bound isolates three complementary intervention points: ϵ sets the starting deviation, ρ_π governs how that deviation propagates, and q_π supplies continuing leakage. When $\rho_\pi < 1$, the initial term decays geometrically toward the limiting allowance $q_\pi/(1-\rho_\pi)$. Because injection control alone leaves ρ_π unconstrained, CRCO penalizes both.

Realizability and tightness. Zero leakage is equivalent to representative-independent quotient dynamics. The affine recurrence is sharp for general kernels, and it is realizable through the native agent–environment factorization when $\epsilon \ge (1-\alpha)q$ (Propositions 11 and 12). Persistent-disturbance and return bounds follow from the same recurrence; Appendix D gives the constructions and extensions.

3.1 Correspondence information and objective control

The affine hinge retains current–successor correspondence that distance marginals discard.

Proposition 1 (Information supplied by state correspondence)**.** *Let $h(d, s) = [s - \bar{\rho} d - \bar{q}]_+$, with $\bar{\rho} > 0$ and $\bar{q} \ge 0$. For $d_\ell \le d_h$ and $s_\ell \le s_h$,*

$$h(d_\ell, s_\ell) + h(d_h, s_h) \le h(d_\ell, s_h) + h(d_h, s_\ell). \tag{11}$$

There exist two pair laws with identical current-distance and successor-distance marginals, and therefore identical absolute and constant-hinge penalties, but strictly different affine-hinge penalties.

Appendix D.2 proves the inequality and gives a strict two-pair construction. Task reward determines which recoveries are useful.

The next result connects that discrimination to task sensitivity without identifying a fitted slope with a uniform contraction constant. Define $V_j = [\widetilde{D}_{j+1} - \bar{\rho}\widetilde{D}_j - \bar{q}]_+$ and the pathwise envelope

$$B_H = \bar{\rho}^H \widetilde{D}_0 + \sum_{j=0}^{H-1} \bar{\rho}^{H-1-j}(\bar{q} + V_j). \tag{12}$$

Theorem 2 (From the training hinge to task sensitivity). *Fix $0 < \bar{\rho} < 1$, $\alpha < 1$, a policy and its marginal-preserving executable coupling. Suppose $\widetilde{D}_j \in [0,1]$ and the normalized endpoint score $g(Z_{H+1}) \in [0,1]$ is L_g-Lipschitz under d_Z. For the clean and perturbed score expectations C_H, P_H under the same initial task law,*

$$|C_H - P_H| \le \min\left\{1, \frac{L_g}{1-\alpha}\mathbb{E}\min\{1, B_H\}\right\}. \tag{13}$$

For M fixed candidate policies independent of the validation sample, let $\widehat{b}_{m,H}$ average $\min\{1, B_H\}$ over n independent paired trajectories per candidate. With probability at least $1-\zeta$, Equation 13 holds for every candidate and every $1 \le H \le K$ after replacing the expectation by

$$\min\left\{1, \widehat{b}_{m,H} + \sqrt{\frac{\log(MK/\zeta)}{2n}}\right\}. \tag{14}$$

The proof combines the pathwise hinge recurrence, the factor $1-\alpha$ of the action-effect metric, and one bounded-mean concentration step per candidate and horizon; it requires neither transport optimality nor independence between rollout steps. Binary success admits $L_g = 1/m_g$ when different outcome labels have semantic separation at least $m_g > 0$. The theorem thus links the optimized excess growth directly to score sensitivity over the measured horizon, and the stopped recurrence accounts for exits when the propagation assumptions hold only inside a recoverable domain.

3.2 Environment–policy composition

The loop factors through the pre-decision and post-decision spaces. Composing the profiles of its two factors controls propagation from either stage and thereby links the action-effect training signal to the direct state-dynamics evaluation used in the experiments.

Assumption 2 (Environment and policy-lift profiles). *Uniformly on the reachable sets, the environment kernel G has affine leakage/gain (q_G, L_G) from d_Y to d_Z, and the policy lift S_π has (q_S, L_S) from d_Z to d_Y. Equations 25–26 state these two bounds explicitly.*

Theorem 3 (Typed kernel-composition profile). *Under Assumption 2, $K_\pi = GS_\pi$ satisfies*

$$q_\pi^0 \le q_S + L_S q_G, \qquad \rho_\pi(q_S + L_S q_G) \le L_S L_G. \tag{15}$$

The pre-decision kernel $H_\pi = S_\pi G$ has affine leakage $q_G + L_G q_S$ and gain $L_G L_S$ under d_Z (Equation 27).

The native metric fixes $q_G = 0$ and admits $L_G \le (1-\alpha)^{-1}$ (Proposition 10). Environment interventions therefore use a common frozen geometry. The product bound is attained under aligned maximizing pairs and optimal composed coupling (Proposition 3); finite-kernel experiments test aligned and misaligned cases. Appendix D.3 resolves policy-lift gain into state and action contributions.

3.3 First exits and terminal outcomes

For distinct absorbing outcomes separated by $d > 0$, any uniform profile satisfies $q \ge (1-\rho)d$, so the affine radius retains persistent outcome separation and local propagation and exit consequences must be accounted for separately. For a recoverable domain, Proposition 8 bounds full-horizon distance by in-domain propagation plus the probability-weighted consequences of each first exit. Successful completion and unresolved exit are separate competing events, and every failed path contributes to the bound. This decomposition dictates the empirical measurements: fixed-pair state dynamics assess local amplification, first-exit frequencies assess departure from the domain, and terminal distances assess post-exit consequences. Enumerable kernels calibrate supremum estimation, while split-confirmed witnesses test expanding pairs directly (Propositions 14 and 16).

4 COUPLED-ROLLOUT CONTRACTION OPTIMIZATION

CRCO converts the affine envelope into an injection penalty and a state-conditioned propagation penalty. Training requires checkpointable or replayable environments and either predicate-based or observation-trained effect measurements; deployment uses one ordinary rollout.

Effect of state conditioning. Consider two rollout steps with the same successor distance, 0.12, but current distances 0.20 and 0.05. With $\bar{\rho} = 0.65$ and $\bar{q} = 0.01$, their affine allowances are 0.14 and 0.0425, producing hinge costs of 0 and 0.0775. The first step has recovered within its allowance, whereas the second has expanded beyond it. An absolute penalty assigns both steps the same cost of 0.12 and therefore erases this distinction.

4.1 COUPLED ROLLOUTS AND TRAINING OBJECTIVE

We sample a task and an intervention time τ from a frozen, training-only reference-prefix pool, clone the environment and controller states, and give the two branches o_τ and an admissible $T(o_\tau)$, respectively. The current policy generates both continuations. Its categorical next-token laws are sampled with exact maximal coupling: shared probability mass emits the same token in both branches, and residual mass preserves each marginal. Simulator randomness is then shared conditional on the paired actions. Together, these steps define the executable coupling γ_{MC}. Each $\widetilde{D}_k$ samples its task-effect cost, while offline empirical OT measures the feasible-coupling gap. Appendix E gives the sampling and likelihood details.

For each pair, $\widetilde{D}_0$ evaluates Equation 4 on the initial actions and their coupled successors; on equivalent inputs, its expectation upper-bounds policy-lift injection. The pair reward is

$$R_{\mathrm{pair}} = \frac{R + R'}{2} - \lambda_{\mathrm{inv}} \widetilde{D}_0 - \lambda_{\mathrm{ctr}} \sum_{k=0}^{K-1} \omega_k \left[\widetilde{D}_{k+1} - \bar{\rho} \widetilde{D}_k - \bar{q} \right]_+ , \tag{16}$$

with $\omega_k = 1/K$. The first penalty limits injected deviation, the affine hinge penalizes only propagation in excess of the state-conditioned allowance, and task return preserves useful execution. Development folds fix $\bar{q}$ and all coefficients, which are shared across matched objectives. Held-out hinges bound mean and CVaR distance under the executable coupling (Propositions 15 and 16).

Proposition 1 explains why each current distance must remain paired with its own successor, and Theorem 2, via the uniform-weight bound in Equation 22, bounds task sensitivity by the same positive violations optimized here; independent OT and first-exit measurements check transport and irreversible outcomes separately from the training reward.

4.2 ACTION-EFFECT MEASUREMENT

ALFWorld and ScienceWorld use symbolic predicates; WebShop uses attribute progress, product identity, and page/action state; τ-bench uses database deltas, user constraints, and irreversible operations. Normalized components are frozen before evaluation. Actions are executed from cloned states and scored by Equation 4, so surface strings carry no cost. Offline OT separately transports action effects at injection and closed-loop successors in the profile. In the observation-trained variant, the same paired rollouts are scored by a frozen encoder learned from training-split observation transitions and rewards, and admissible transformations are restricted to interface symmetries guaranteed by renderer construction. This variant therefore requires no predicate or action-effect oracle, although replayable paired training remains an explicit requirement of the method. Neither metric enters the deployed prompt; Appendix F specifies weights, stochastic-transition estimation, coupling audits, and learned or human alternatives.

4.3 OPTIMIZATION AND MATCHED CONTROLS

Gradient path. Each four-pair group shares one task and reference checkpoint. Frozen semantic extractors produce detached rewards; gradients pass through the maximal-coupling joint likelihood of all return-relevant continuation tokens, with one clipped minibatch epoch per fresh rollout batch (Appendix E; Equations 68–70).

We instantiate CRCO with GRPO (Shao et al., 2024) and standardize advantages within each task–checkpoint group. Every objective shares initialization, environment access, branch count, budgets, optimizer, coupling, and injection penalty, so differences between objectives isolate the propagation functional. The primary control, matched absolute propagation, changes only that functional:

$$R^{\mathrm{abs}}_{\mathrm{pair}} = \frac{R + R'}{2} - \lambda_{\mathrm{inv}} \widetilde{D}_0 - \lambda_{\mathrm{abs}} \sum_{k=0}^{K-1} \omega_k \widetilde{D}_{k+1}. \tag{17}$$

Paired GRPO removes both penalties; one-step consistency retains only the injection penalty; progress shaping and terminal divergence provide reward- and endpoint-based controls. Constant-threshold and shuffled-current hinges match activation and mean penalty while removing threshold variability and current–successor correspondence, respectively. A direct-state hinge tests the same principle without look-ahead overlap, and product-coupling variants separate coupling entropy. Appendix M gives definitions and coefficient matching.

Relation to existing objectives. Input regularizers, one-step consistency, and trajectory divergences target complementary failure modes. CRCO conditions the successor allowance on current separation; the constant and shuffled-current hinges isolate this ingredient. Appendix R discusses prior work.

5 Experiments

The experiments form the third contribution and test the first two. **RQ1** tests transfer across backbones and environments; **RQ2** tests whether the gain arises from state-conditioned propagation control (Contribution 2); and **RQ3** tests whether short profiles predict long-horizon outcomes and reduce selection cost (Contribution 1). Published external anchors are attributed separately in Appendix Q.

5.1 Experimental scope

The design crosses nine backbones with eight interactive settings under two execution protocols. ALFWorld (Shridhar et al., 2020), ScienceWorld (Wang et al., 2022), WebShop (Yao et al., 2022a), and the original τ-bench (Yao et al., 2024) use 1,099 disjoint reference-prefix test clusters, whereas WebArena (Zhou et al., 2023), WorkArena++ L2 (Boisvert et al., 2024), τ^2-bench telecom (Barres et al., 2025), and τ-Knowledge (Shi et al., 2026) start from reset and run to termination. The two protocols define distinct estimands and are never pooled. The nine backbones are Qwen2.5 (Qwen Team, 2024), Gemma 3 (Gemma Team et al., 2025), Llama 3.1 (Grattafiori et al., 2024), Qwen3.5/Qwen3.6/Qwen3.8 (Qwen Team, 2026a;b;c), and Gemma 4 (Google DeepMind, 2026); comparisons remain within a backbone. Table 1 aligns six objectives on every backbone, and Table 2 reports the model–setting cells as two 36-cell blocks with a paired-seed SD for each gain. The anchor protocol uses eight Qwen2.5-7B and six Gemma-3-12B seed blocks, and every other backbone uses six (Appendices C and P).

The baseline ladder separates training strength from mechanism. SFT (Ouyang et al., 2022), PPO (Schulman et al., 2017), GRPO (Shao et al., 2024), HCAPO (Tan et al., 2026), canonicalization, observation augmentation (Kostrikov et al., 2020), and the study-specific ToolRL-DR control cover learning, credit assignment, preprocessing, and input randomization. Paired GRPO and one-step consistency isolate pairing and injection control; matched absolute, activation-matched constant, and shuffled-current objectives remove state conditioning, selective activation, and current–successor correspondence, respectively. Tail-divergence, worst-family DRO, and trajectory-MMD are resource-matched robust alternatives defined in Appendix M.

5.2 RQ1: Robustness across models and benchmarks

Under the reference-prefix protocol, CRCO has the highest perturbed-score point estimate among the six matched objectives on all nine backbones in Table 1, exceeding matched absolute propagation by 0.5–5.2 P points across dense and MoE backbones. Over eight settings (Table 2), the reference-prefix and reset-to-terminal blocks have nine and eight positive backbone means, with equal-cell mean gains of $+3.2$ and $+2.5$ points. The advantage is consistent at the protocol level but not uniform across cells: gains range from -0.6 to $+7.8$ under reference prefixes and from -1.5 to $+7.0$ after reset, and

Table 1: Matched-objective comparison on nine backbones under the reference-prefix protocol. Entries are macro perturbed scores P over ALFWorld, ScienceWorld, WebShop, and τ-bench (higher is better); the CRCO column also reports clean score C. G/I/A/C/S: paired GRPO, one-step consistency, matched absolute, activation-matched constant, and shuffled-current. ΔP: CRCO minus A. Bold marks the rows where CRCO has the highest P; this holds on all nine backbones, with ΔP ranging from $+0.5$ to $+5.2$.

Backbone	G	I	A	C	S	CRCO C/P ↑	ΔP vs. A
Qwen2.5-7B-Instruct	42.0	47.3	50.7	51.2	50.9	**61.6 / 53.3**	**+2.6**
Gemma-3-12B-IT	44.8	48.6	52.4	53.1	52.0	**63.0 / 55.7**	**+3.3**
Qwen2.5-14B	47.0	51.5	54.5	55.6	56.2	**65.7 / 59.7**	**+5.2**
Llama-3.1-8B	39.4	45.6	48.2	48.3	49.0	**59.4 / 50.2**	**+2.0**
Qwen3.5-9B	53.0	58.7	62.8	62.6	62.3	**70.6 / 67.6**	**+4.8**
Qwen3.6-35B-A3B	59.7	64.9	69.9	68.2	68.0	**75.6 / 71.4**	**+1.5**
Qwen3.8-27B	73.0	72.4	75.2	75.5	75.6	**81.7 / 80.2**	**+5.0**
Gemma-4-26B-A4B-IT	55.2	60.4	68.0	64.1	63.7	**73.1 / 68.5**	**+0.5**
Gemma-4-31B-IT	65.4	69.5	73.3	73.1	73.5	**80.2 / 77.4**	**+4.1**

Table 2: Nine-backbone CRCO matrix under two separate estimands. Each cell gives CRCO perturbed score P on the upper line and its gain over matched absolute propagation $\pm$ paired-seed SD on the lower line. SD uses eight seed blocks for Qwen2.5-7B and six elsewhere. Block A uses reference-prefix and block B reset-to-terminal execution; the blocks are not pooled. Equal-cell mean gains are $+3.2$ (A) and $+2.5$ (B), with positive backbone means on nine backbones in A and eight in B.

	A. Reference-prefix					B. Reset-to-terminal		
Model	ALF	SciW	WebS	τ-b	WebA	Work++	τ^2	τ-K
Q2.5-7B	66.0 +2.3 ± 1.3	41.0 +3.5 ± 2.1	62.5 +1.5 ± 0.8	43.7 +3.1 ± 1.6	22.5 +5.4 ± 1.8	9.7 -0.1 ± 1.2	40.9 -0.8 ± 1.7	11.5 +4.8 ± 0.7
G3-12B	67.4 +1.5 ± 0.7	45.6 +6.8 ± 2.2	64.0 -0.2 ± 1.3	45.8 +5.1 ± 1.9	22.7 -0.9 ± 2.1	14.8 +5.7 ± 1.5	44.6 +4.3 ± 2.0	10.1 -0.3 ± 0.9
Q2.5-14B	76.2 +7.4 ± 0.6	46.0 +3.8 ± 1.7	69.5 +4.6 ± 1.1	47.1 +5.0 ± 1.5	29.8 +6.0 ± 1.7	13.6 -1.0 ± 1.1	50.5 +5.2 ± 1.6	13.1 +2.7 ± 0.8
L3.1-8B	61.6 +0.8 ± 1.3	38.1 +3.7 ± 2.5	57.3 -0.4 ± 1.6	43.8 +3.9 ± 1.9	17.8 -1.1 ± 2.2	12.0 +5.7 ± 0.9	39.7 +5.9 ± 2.0	7.3 -0.4 ± 0.6
Q3.5-9B	79.3 -0.2 ± 0.5	58.7 +7.1 ± 1.8	74.1 +4.8 ± 0.9	58.3 +7.5 ± 1.6	29.6 +3.1 ± 1.7	18.7 -1.2 ± 1.4	64.6 +6.2 ± 1.5	15.9 +2.7 ± 0.8
Q3.6-35B	83.4 -0.6 ± 0.4	63.6 +5.8 ± 1.5	75.4 +0.0 ± 0.7	63.2 +0.8 ± 1.4	43.7 -1.4 ± 1.9	28.7 +6.3 ± 1.6	66.9 -1.5 ± 1.3	24.2 +5.5 ± 0.7
Q3.8-27B	88.1 +0.6 ± 0.4	74.3 +7.5 ± 1.2	83.0 +4.1 ± 0.7	75.4 +7.8 ± 1.1	55.1 +7.0 ± 1.4	30.6 +1.4 ± 1.1	71.2 +4.4 ± 1.0	27.3 +4.7 ± 0.6
G4-26B	83.1 +0.5 ± 0.9	59.1 -0.5 ± 2.3	74.1 +1.1 ± 1.4	57.6 +0.8 ± 2.0	38.8 -0.9 ± 1.9	24.5 -1.1 ± 1.5	61.2 +0.8 ± 1.5	19.2 -0.6 ± 0.7
G4-31B	89.1 +4.8 ± 0.5	69.9 +5.2 ± 1.6	78.8 -0.3 ± 0.8	71.6 +6.6 ± 1.5	48.3 +3.3 ± 1.6	30.9 +5.1 ± 1.2	65.1 +4.2 ± 1.1	25.2 +0.3 ± 0.6

paired-seed SDs separate small local effects from the protocol-level advantage. WorkArena++ L2 and τ-Knowledge are the hardest settings, and the largest gains do not follow a fixed column pattern. Appendices O, M, and Q report full C/P cells, baselines, and external anchors.

5.3 RQ2: Ablations and propagation dynamics

On the two instrumented anchors, Table 3 gives predicate-metric CRCO a 3.0-point mean P advantage over matched absolute propagation while reducing held-pair direct-state OT gain by 0.18, so the score gain coincides with an independently measured reduction in propagation. The observation-trained metric retains a 2.6-point advantage and a 0.14 gain reduction without latent predicates or

Table 3: Matched ablation on two instrumented anchors. P is perturbed score; $\widehat{g}_A$ is held-pair direct-state OT gain. The final columns average effects relative to matched absolute (A) over both anchors. The observation-trained metric uses retained observations, actions, rewards, and next observations only. Arrows mark the preferred direction. Only state-conditioned CRCO gains ≥ 2.6 P and cuts $\widehat{g}_A$ by ≥ 0.14.

	Qwen2.5-7B		Gemma-3-12B		Mean effect vs. A	
Objective	$P \uparrow$	$\widehat{g}_A \downarrow$	$P \uparrow$	$\widehat{g}_A \downarrow$	$\Delta P \uparrow$	$\Delta\widehat{g}_A \downarrow$
Paired GRPO	42.0	1.08	44.8	1.14	-8.2	+0.21
One-step consistency	47.3	1.04	48.6	0.98	-3.6	+0.11
Matched absolute	50.7	0.86	52.4	0.94	+0.0	+0.00
Activation-matched constant hinge	51.2	0.83	53.1	0.87	+0.6	-0.05
Shuffled-current hinge	50.9	0.85	52.0	0.91	-0.1	-0.02
CRCO (obs.-metric)	53.0	0.74	55.3	0.78	+2.6	-0.14
CRCO	**53.3**	**0.71**	**55.7**	**0.74**	**+3.0**	**-0.18**

action-consequence oracles. The activation-matched constant ($+0.6$) and shuffled-current (-0.1) hinges keep the hinge form but remove state conditioning and recover little of this benefit, so current–successor correspondence rather than selective activation accounts for most of the gain (Proposition 1). Independent draws, direct-state training, graded reassignment, recovery interventions, and metric triangulation provide complementary tests (Appendices A, G.1, H, and I).

5.4 RQ3: Long-horizon prediction and selection

Short profiles predict long-horizon outcomes: across 112 matched checkpoint pairs, the success gap between low- and high-gain checkpoints grows from 0.2 points at horizon 1 to 12.8 points at horizon 24, and profiles reach 0.863/0.816 AUROC under method/benchmark holdout. This validity also lowers selection cost: across 28 held-out model–benchmark–seed panels and 224 candidates, five-step selection matches the test oracle 23 times with 0.34-point regret, versus 21 matches and 0.48-point regret for full-horizon validation, with 40.7% fewer robustness calls (Appendices A and J).

6 Conclusion

For measurement, the affine Wasserstein profile separates semantic leakage from feedback amplification in a policy-independent task geometry and links both to outcomes through composition and first-exit bounds. For optimization, CRCO conditions each successor separation on the current deviation under marginal-preserving coupling while retaining task reward; Proposition 1 and Theorem 2 establish why this correspondence is informative and how the optimized hinge bounds score sensitivity. Empirically, CRCO has the highest matched perturbed-score point estimate on all nine backbones; the separately analyzed reference-prefix and reset-to-terminal blocks yield nine and eight positive backbone means and equal-cell gains of 3.2 and 2.5 points; reassignment, held-pair, recovery, and five-step analyses establish the mechanism, its outcome effect, and selection efficiency; and the observation-trained metric retains a 2.6-point mean gain without latent predicates.

Implications. State correspondence makes recovery an optimization target: the same successor separation receives different penalties depending on the preceding deviation. The matched constant and shuffled-current controls isolate this information, while independent direct-state measurements connect it to reduced propagation. The observation-trained metric extends this result beyond privileged predicates, using task-effect information from observations and rewards. Five-step profiles also turn propagation measurements into checkpoint decisions, concentrating paired evaluation in development while deployment remains one ordinary rollout. These results connect feedback measurement, training, and checkpoint selection. Training requires replayable states, and the theoretical guarantees retain their stated metric and domain assumptions.

AI USE STATEMENT

We used LLMs to edit and draft manuscript text, summarize and interpret reported results, suggest organization, locate references, and revise LaTeX tables, figure presentation, and document formatting. AI assistance also supported editing of table-generation and plotting scripts and reformatting of result files. The authors are responsible for verifying all claims, citations, numerical results, code, and AI-assisted text against the underlying research records, and for the final submission.

ETHICS STATEMENT

Training agents to recover from benign observation variation may reduce accidental failures, but the same persistence could also let an agent complete harmful tasks more reliably. Our evaluation therefore uses sandboxed simulators and excludes irreversible external actions. Because contraction is certified only in the chosen geometry, effect-grounded distances should cover all safety-relevant changes rather than task progress alone; otherwise a policy can appear contractive while consistently causing unmeasured harm.

REPRODUCIBILITY STATEMENT

The main text specifies the learning objective, identifying comparisons, and statistical estimands, while the appendix documents the proofs, implementation details, and evaluation definitions. The anonymous supplementary artifact provides an executable reference implementation of the CRCO core, including task-semantic geometry, maximal action coupling, paired-rollout objects, affine state-conditioned violations, absolute and constant-threshold controls, and return construction. Its modules support correspondence checks with identical distance marginals, action-effect transition costs, injection and propagation penalties, and deterministic numerical validation. The package also includes the equations, API descriptions, data inventory, and numerical summaries needed to inspect the implementation and connect each computation to the reported analysis. The reference functions use ordinary Python scalars and tuples, so they can be embedded directly in simulator loops, policy-training code, or independent analysis programs.

REFERENCES

David Angeli. A lyapunov approach to incremental stability properties. *IEEE Transactions on Automatic Control*, 47(3):410–421, 2002. doi: 10.1109/9.989067.

David Angeli. Further results on incremental input-to-state stability. *IEEE Transactions on Automatic Control*, 54(6):1386–1391, 2009. doi: 10.1109/TAC.2009.2015561.

Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. τ^2-bench: Evaluating conversational agents in a dual-control environment, 2025. URL https://arxiv.org/abs/2506.07982.

Léo Boisvert, Megh Thakkar, Maxime Gasse, Massimo Caccia, Thibault Le Sellier De Chezelles, Quentin Cappart, Nicolas Chapados, Alexandre Lacoste, and Alexandre Drouin. Workarena++: Towards compositional planning and reasoning-based common knowledge work tasks, 2024. URL https://arxiv.org/abs/2407.05291.

Peter Buchholz. Exact and ordinary lumpability in finite markov chains. *Journal of Applied Probability*, 31(1):59–75, 1994. doi: 10.2307/3215235.

Yinlam Chow, Ofir Nachum, Edgar Duenez-Guzman, and Mohammad Ghavamzadeh. A lyapunov-based approach to safe reinforcement learning. *arXiv preprint arXiv:1805.07708*, 2018.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transportation distances. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

Thibault Le Sellier De Chezelles, Maxime Gasse, Alexandre Drouin, Massimo Caccia, et al. The browsergym ecosystem for web agent research. *arXiv preprint arXiv:2412.05467*, 2024.

Alexandre Drouin, Maxime Gasse, Massimo Caccia, Issam H. Laradji, Manuel Del Verme, Tom Marty, Léo Boisvert, Megh Thakkar, Quentin Cappart, David Vazquez, Nicolas Chapados, and Alexandre Lacoste. Workarena: How capable are web agents at solving common knowledge work tasks? *arXiv preprint arXiv:2403.07718*, 2024.

Norm Ferns, Prakash Panangaden, and Doina Precup. Bisimulation metrics for continuous markov decision processes. *SIAM Journal on Computing*, 40(6):1662–1714, 2011. doi: 10.1137/10080484X.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. doi: 10.48550/arXiv.2503.19786.

Robert Givan, Thomas Dean, and Matthew Greig. Equivalence notions and model minimization in markov decision processes. *Artificial Intelligence*, 147(1–2):163–223, 2003. doi: 10.1016/S0004-3702(02)00376-4.

Google DeepMind. Gemma 4: Model card and official checkpoints, 2026. URL `https://ai.google.dev/gemma/docs/core/model_card_4`. Model family and 26B-A4B/31B instruction-tuned variants verified in the official google-deepmind/gemma repository, accessed 2026-09-23.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. Critic: Large language models can self-correct with tool-interactive critiquing. *arXiv preprint arXiv:2305.11738*, 2023.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Sch"olkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

Bhaskar Gurram. Auditing automated evaluation, error propagation, and runtime mitigation in tool-using language agents. *arXiv preprint arXiv:2604.16706*, 2026.

Garud N. Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2):257–280, 2005. doi: 10.1287/moor.1040.0129.

Ilya Kostrikov, Denis Yarats, and Rob Fergus. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. *arXiv preprint arXiv:2004.13649*, 2020.

Xingzuo Li, Kehai Chen, Yunfei Long, Xuefeng Bai, Yong Xu, and Min Zhang. Generator-assistant stepwise rollback framework for large language model agent. *arXiv preprint arXiv:2503.02519*, 2025.

Qiao Liang, Yuke Zhu, Chao Ge, Lei Yang, Ying Shen, Bo Zheng, and Sheng Guo. Learning from the irrecoverable: Error-localized policy optimization for tool-integrated llm reasoning. *arXiv preprint arXiv:2602.09598*, 2026.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, et al. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688*, 2023.

Winfried Lohmiller and Jean-Jacques E. Slotine. On contraction analysis for non-linear systems. *Automatica*, 34(6):683–696, 1998. doi: 10.1016/S0005-1098(98)00019-3.

Xufang Luo, Yuge Zhang, Zhiyuan He, Zilong Wang, Siyun Zhao, Dongsheng Li, Luna K. Qiu, and Yuqing Yang. Agent lightning: Train any ai agents with reinforcement learning. *arXiv preprint arXiv:2508.03680*, 2025.

Chang Ma, Junlei Zhang, Zhihao Zhu, Cheng Yang, Yujiu Yang, Yaohui Jin, Zhenzhong Lan, Lingpeng Kong, and Junxian He. Agentboard: An analytical evaluation board of multi-turn llm agents. *arXiv preprint arXiv:2401.13178*, 2024.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*, 2023.

Arnab Nilim and Laurent El Ghaoui. Robust control of markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005. doi: 10.1287/opre.1050.0216.

Yann Ollivier. Ricci curvature of markov chains on metric spaces. *Journal of Functional Analysis*, 256(3):810–864, 2009. doi: 10.1016/j.jfa.2008.11.001.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

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

Qwen Team. Qwen3.6-35B-A3B: Agentic coding power, now open to all, April 2026b. URL `https://qwen.ai/blog?id=qwen3.6-35b-a3b`.

Qwen Team. Qwen3.8-27B model card, August 2026c. URL `https://huggingface.co/Qwen/Qwen3.8-27B`.

R. Tyrrell Rockafellar and Stanislav Uryasev. Optimization of conditional value-at-risk. *The Journal of Risk*, 2(3):21–41, 2000. doi: 10.21314/jor.2000.038.

Daniel Rudolf and Nikolaus Schweizer. Perturbation theory for markov chains via wasserstein distance. *Bernoulli*, 24(4A):2610–2639, 2018. doi: 10.3150/17-BEJ938.

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

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

Quan Shi, Alexandra Zytek, Pedram Razavi, Karthik Narasimhan, and Victor Barres. τ-knowledge: Evaluating conversational agents over unstructured knowledge. *arXiv preprint arXiv:2603.04370*, 2026.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *arXiv preprint arXiv:2303.11366*, 2023.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Cote, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. Alfworld: Aligning text and embodied environments for interactive learning. *arXiv preprint arXiv:2010.03768*, 2020.

Haotian Sun, Yuchen Zhuang, Lingkai Kong, Bo Dai, and Chao Zhang. Adaplanner: Adaptive planning from feedback with language models. *arXiv preprint arXiv:2305.16653*, 2023.

Hui-Ze Tan, Xiao-Wen Yang, Hao Chen, Jie-Jing Shao, Yi Wen, Yuteng Shen, Weihong Luo, Xiku Du, Lan-Zhe Guo, and Yu-Feng Li. Hindsight credit assignment for long-horizon llm agents. *arXiv preprint arXiv:2603.08754*, 2026.

Ruoyao Wang, Peter Jansen, Marc-Alexandre Cote, and Prithviraj Ammanabrolu. Scienceworld: Is your agent smarter than a 5th grader? *arXiv preprint arXiv:2203.07540*, 2022.

Zixuan Wang, Yuchen Yan, Hongxing Li, Teng Pan, Dingming Li, Ruiqing Zhang, Weiming Lu, Jun Xiao, Yueting Zhuang, and Yongliang Shen. Milestone-guided policy learning for long-horizon language agents. *arXiv preprint arXiv:2605.06078*, 2026.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. Webshop: Towards scalable real-world web interaction with grounded language agents. *arXiv preprint arXiv:2207.01206*, 2022a.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022b.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. tau-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*, 2024.

Siyu Yuan, Zehui Chen, Zhiheng Xi, Junjie Ye, Zhengyin Du, and Jiecao Chen. Agent-r: Training language model agents to reflect via iterative self-training. *arXiv preprint arXiv:2501.11425*, 2025.

Amy Zhang, Rowan McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. In *International Conference on Learning Representations*, 2021.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*, 2023.

Kunlun Zhu, Zijia Liu, Bingxuan Li, Muxin Tian, Yingxuan Yang, Jiaxun Zhang, Pengrui Han, Qipeng Xie, Fuyang Cui, Weijia Zhang, Xiaoteng Ma, Xiaodong Yu, Gowtham Ramesh, Jialian Wu, Zicheng Liu, Pan Lu, James Zou, and Jiaxuan You. Where llm agents fail and how they can learn from failures. *arXiv preprint arXiv:2509.25370*, 2025.

Yu Zhuang, Kefei Chen, Yitong Duan, Shuxin Zheng, Jian Li, and Xu-Yao Zhang. Agentrewind: Recoverable execution for long-horizon llm agents. *arXiv preprint arXiv:2608.14380*, 2026.

APPENDIX GUIDE

The supplement follows the paper's claim chain. Appendix A summarizes the main-text experiments; Appendices B–G fix the protocol, mathematical objects, algorithm, and metrics; Appendices H–J test mechanism, recovery, prediction, and checkpoint selection; and Appendices K–O report transfer and stress tests. Statistics, external anchors, related work, and the disclosure statements close the supplement. Each block states its estimand, sampling unit, and matched comparison so that every main-text claim can be traced to its supporting record. The ordering keeps each claim adjacent to its audit.

A MAIN EXPERIMENT DETAILS

This appendix gives the figures and expanded tables behind RQ1–RQ3. They share the main text's estimands and task partitions; repeated rows are diagnostic views of the same model–benchmark matrix. Captions identify the averaging unit and held-out split, while figures provide visual summaries of the same matched evaluations. All comparisons retain the same checkpoint and task pairing.

Table 4: Fully instrumented matched-objective diagnostics on the two anchor backbones; within each backbone, all objectives share the paired sampling protocol and optimization budgets. C and P are macro-averaged clean and perturbed task scores. $\widehat{q}^{\,\mathrm{zero}}_{0.95}$ is the zero-class leakage quantile, $\widehat{b}_1$ the adjacent-cost slope, $\widehat{g}_A$ the direct-state OT gain, and $\widetilde{D}_0$ the initial coupled distance, with lower values better for all four diagnostics; $H(A)$ denotes action entropy. On both anchors, CRCO attains the highest C and P and the lowest leakage quantile, adjacent-cost slope, and direct-state OT gain; for initial coupled distance, it ties for the lowest value on the Qwen anchor, whereas one-step consistency is lowest on the Gemma anchor. Table 1 gives the all-backbone performance matrix.

	Task performance		Propagation diagnostics				Policy
Objective	C ↑	P ↑	$\widehat{q}^{\,\mathrm{zero}}_{0.95}$ ↓	$\widehat{b}_1$ ↓	$\widehat{g}_A$ ↓	$\widetilde{D}_0$ ↓	$H(A)$
Qwen2.5-7B-Instruct (8 seeds)							
Paired GRPO	60.5	42.0	0.040	1.13	1.08	0.121	1.45
One-step consistency	59.7	47.3	0.014	1.10	1.04	0.116	1.43
Matched absolute propagation	60.0	50.7	0.014	0.81	0.86	0.115	1.42
Activation-matched constant hinge	60.4	51.2	0.013	0.80	0.83	**0.114**	1.42
Shuffled-current hinge	60.2	50.9	0.014	0.84	0.85	0.115	1.43
CRCO (affine propagation)	**61.6**	**53.3**	**0.012**	**0.69**	**0.71**	**0.114**	1.42
Gemma-3-12B-IT (6 seeds)							
Paired GRPO	62.1	44.8	0.035	1.21	1.14	0.128	1.34
One-step consistency	61.0	48.6	0.013	1.05	0.98	**0.103**	1.30
Matched absolute propagation	61.8	52.4	0.016	0.91	0.94	0.110	1.32
Activation-matched constant hinge	62.4	53.1	0.012	0.78	0.87	0.107	1.28
Shuffled-current hinge	61.7	52.0	0.015	0.89	0.91	0.112	1.33
CRCO (affine propagation)	**63.0**	**55.7**	**0.009**	**0.63**	**0.74**	0.104	1.26

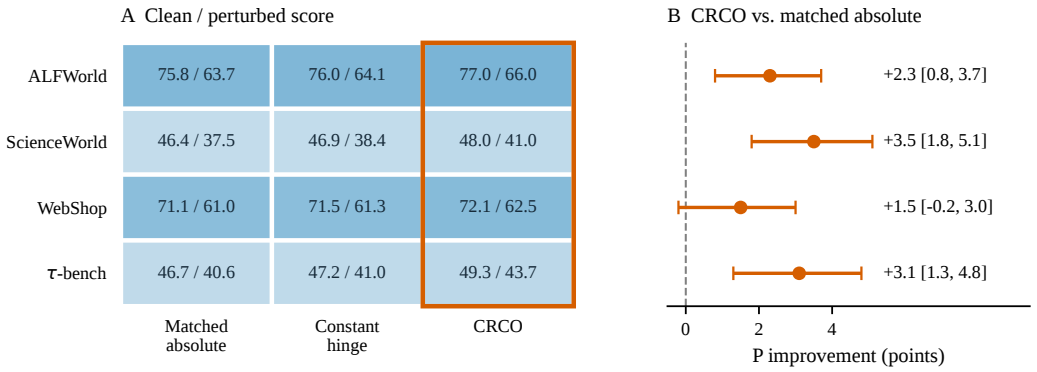


Figure 2: Task-level clean and perturbed performance under matched objectives. **A:** Each cell reports C / P for Qwen2.5-7B-Instruct; shading encodes P on a common 0–100 scale, and the outline marks CRCO. **B:** Paired P differences relative to matched absolute propagation, shown with the reported intervals.

Table 5: Evaluation ledger for the transfer, mechanism, prediction, selection, and stress evidence. Counts give the design scale and independent unit behind each evidence block; overlapping diagnostic views are not added as independent replications.

Evidence block	Design scale and independent unit	Primary endpoints
Model–task transfer	9 backbones, 8 settings, 72 model–setting cells; 8 Qwen anchor seeds and 6 seeds elsewhere	Clean/perturbed score, within-backbone matched gain
Mechanism identification	2 instrumented anchors, 4 original tasks, matched thresholds and rollout budgets	Leakage, adjacent-cost slope, held-pair OT gain, recovery and first exit
Prediction	112 matched checkpoint pairs plus 80 randomized target-sweep policies in 16 blocks	Horizon-24 gap, AUROC, calibration, recurrence coverage
Checkpoint selection	28 model–task–seed panels, 8 checkpoints each, 224 candidate decisions	Oracle agreement, test regret, logical slots, recorded calls
Stress and transfer	Held-out transformation families, orders 1–4, persistent noise, tail risk, deployment shifts	Score retention, CVaR, irreversible failure, cross-domain profile stability

Table 6: Reset-to-terminal coverage of three recent backbones across all eight interactive settings. Each candidate generates its own prefix from task reset, and every task remains in the score denominator. A is matched absolute propagation; entries are clean/perturbed scores, and ΔP is CRCO minus A on the perturbed score. CRCO has the higher perturbed score in all but three cells; these rows are reported separately from the reference-prefix results.

Backbone	Benchmark	A C/P	CRCO C/P	ΔP
Qwen3.5-9B	ALFWorld	80.6 / 75.0	81.3 / 79.6	+4.6
	ScienceWorld	56.8 / 50.1	58.0 / 52.3	+2.2
	WebShop	74.9 / 69.3	75.4 / 69.0	-0.3
	τ-bench	56.1 / 48.2	57.5 / 54.0	+5.8
	WebArena	32.3 / 26.5	34.1 / 29.6	+3.1
	WorkArena++ L2	20.1 / 19.9	22.0 / 18.7	-1.2
	τ^2-bench telecom	69.6 / 58.4	72.0 / 64.6	+6.2
	τ-Knowledge banking	16.1 / 13.2	17.8 / 15.9	+2.7
Qwen3.8-27B	ALFWorld	88.9 / 85.7	89.6 / 86.8	+1.1
	ScienceWorld	73.1 / 64.4	74.0 / 70.3	+5.9
	WebShop	81.7 / 77.0	82.1 / 80.0	+3.0
	τ-bench	72.8 / 66.4	74.0 / 69.0	+2.6
	WebArena	56.9 / 48.1	58.7 / 55.1	+7.0
	WorkArena++ L2	33.6 / 29.2	35.1 / 30.6	+1.4
	τ^2-bench telecom	75.2 / 66.8	77.4 / 71.2	+4.4
	τ-Knowledge banking	26.7 / 22.6	28.4 / 27.3	+4.7
Gemma-4-31B-IT	ALFWorld	87.0 / 82.9	88.1 / 86.6	+3.7
	ScienceWorld	70.4 / 62.5	71.7 / 65.9	+3.4
	WebShop	80.8 / 77.3	81.4 / 77.2	-0.1
	τ-bench	69.7 / 61.6	71.2 / 67.7	+6.1
	WebArena	51.8 / 45.0	53.4 / 48.3	+3.3
	WorkArena++ L2	31.8 / 25.8	33.2 / 30.9	+5.1
	τ^2-bench telecom	71.0 / 60.9	73.2 / 65.1	+4.2
	τ-Knowledge banking	25.0 / 24.9	26.9 / 25.2	+0.3

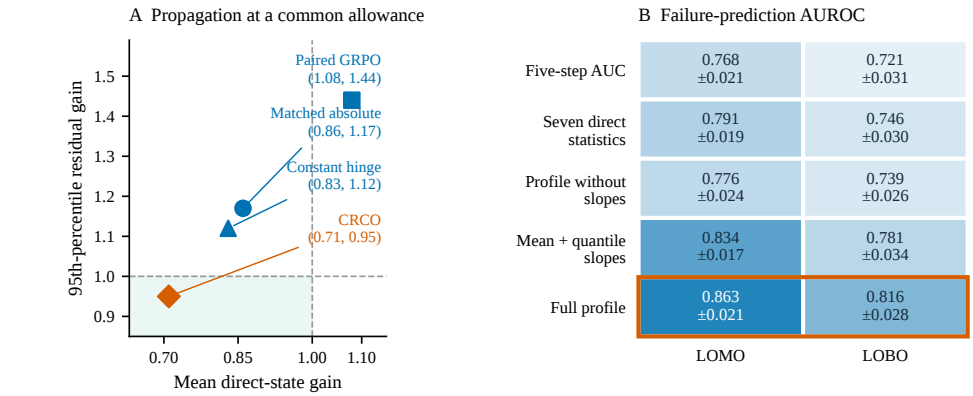


Figure 3: Independently measured propagation and its predictive information. **A:** Mean direct-state OT gain and residual gain at a common leakage allowance. **B:** Five-step failure-prediction AUROC under method-family and benchmark holdouts. The profile and prediction appendices report the complete values.

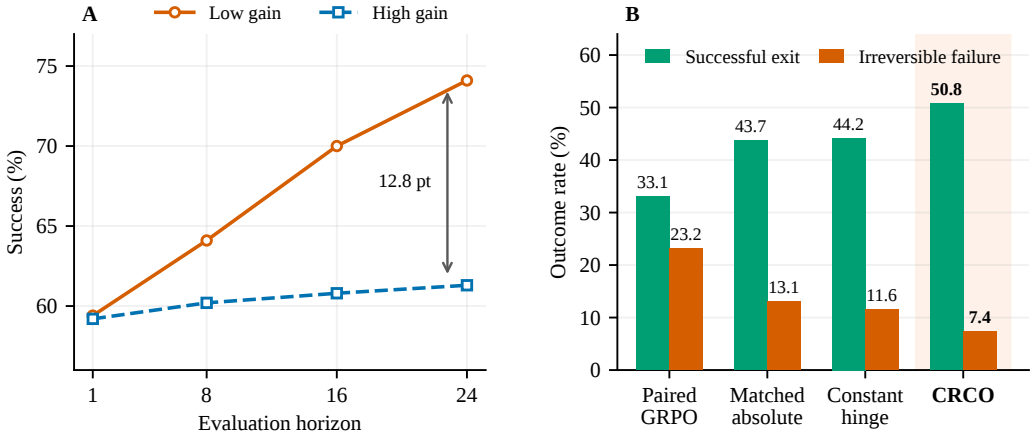


Figure 4: Long-horizon consequences of checkpoint-level propagation gain. **A:** Cumulative success by evaluation horizon for matched low/high-gain checkpoint groups. **B:** First-exit outcomes from the same initial recoverable cohort, with later exits retained.

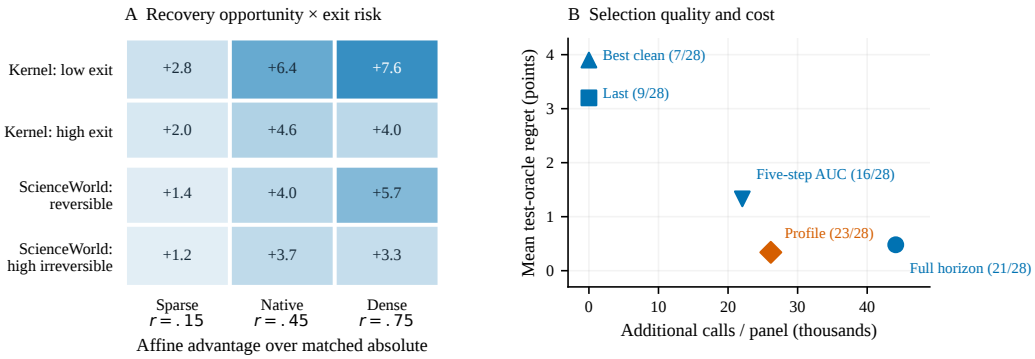


Figure 5: Recovery and selection details. **A:** Outcome differences, affine minus absolute propagation, under matched recovery interventions. **B:** Selection cost and test-oracle regret of checkpoint selectors over held-out direct-state panels.

B NOTATION TABLE

The symbols are organized by the role they play in the analysis: the table below covers the state spaces, the task-semantic geometry, and the theoretical profiles of the measurement theory, and the table on the following page covers coupled-rollout training and the empirical diagnostics. Throughout, primes distinguish the two coupled branches, a superscript $+$ marks a successor, hats mark empirical estimates, and the bars on $(\bar{q}, \bar{\rho})$ mark training targets rather than theoretical quantities.

Theoretical profiles, executable costs, and fitted summaries are distinct objects and should not be read interchangeably.

STATE SPACES, SEMANTIC GEOMETRY, AND THEORETICAL PROFILES

Symbol	Meaning and role
x, h, o, a	Latent task state, retained history/controller state, observation, and action.
$z = (x, h, o) \in \mathcal{Z}$	Augmented pre-decision state; the policy observes (h, o).
$y = (z, a) \in \mathcal{Y}$	Post-decision state, with $\mathcal{Y} = \mathcal{Z} \times \mathcal{A}$.
$\Pi_\pi(\cdot \mid z)$	Action distribution of policy π, conditioned on (h, o).
$G : \mathcal{Y} \to \mathcal{P}(\mathcal{Z})$	Environment/history kernel: execute an action and update observation and retained history; $\mathcal{P}$ denotes probability laws.
$S_\pi : \mathcal{Z} \to \mathcal{P}(\mathcal{Y})$	Policy lift: retain z and append an action sampled from Π_π.
$K_\pi = GS_\pi, H_\pi = S_\pi G$	Post-decision and pre-decision closed-loop kernels, respectively; composition follows execution order.
ϕ_j, δ_j, w_{bj}	Frozen semantic extractor, coordinate pseudometric, and benchmark-specific weight for coordinate j.
d_Z, d_X	Full task-semantic pseudometric; its renormalized restriction to latent-state coordinates.
d_Y, α	Post-decision action-effect pseudometric and its current-state weight (Equation 3).
$\mathcal{W}_d(\mu, \nu)$	Wasserstein distance between laws μ, ν with ground cost d.
$\Gamma(\mu, \nu)$	Set of couplings with marginals μ and ν.
$y \sim y', \mathcal{Y}/\sim$	Zero-distance equivalence, $d_Y(y, y') = 0$, and its semantic quotient.
$\mathcal{Y}_{\mathrm{adm}}$	Reachable admissible set over which the theoretical profile is defined.
$T \in \mathcal{T}, \tau$	Admissible observation transformation and intervention time.
μ_0, ν_0, ϵ	Initial post-decision laws and an upper bound on their Wasserstein separation.
q^0_π	Supremum successor leakage over zero-distance post-decision pairs.
$q, \rho_\pi(q)$	Independently fixed leakage budget and uniform residual amplification at that budget (Definition 1).
(q_π, ρ_π)	Abbreviation for one selected theoretical profile $(q, \rho_\pi(q))$.
$(q_G, L_G), (q_S, L_S)$	Affine leakage/gain bounds for the environment kernel and policy lift (Assumption 2).
$q_{\mathrm{policy}}, L_{\mathrm{policy}}$	Action-distribution leakage and sensitivity under the action-effect cost $c_{z,z'}$.
$L_z, L_a, q_\star$	State/action sensitivity constants and the resulting composite leakage budget (Equation 55).
$\mathcal{R}, \tau_{\mathcal{R}}$	Recoverable pair domain and its first-exit time.
$J, M_{t,j}, F_H$	First-exit cause, conditional terminal-distance bound for exit cell (t, j), and irreversible failure by horizon H.

COUPLED-ROLLOUT TRAINING AND EMPIRICAL DIAGNOSTICS

Symbol	Meaning and role
γ_K, ξ_G	Executable couplings of successor post-decision laws and look-ahead environment laws.
$\gamma = (\gamma_K, \xi_G)$	The two coupling layers defining the executable mean cost.
γ_{MC}	Executable rollout sampler using tokenwise maximal action coupling and shared simulator randomness.
$\mathcal{W}(y, y')$	Theoretical Wasserstein response between $K_\pi(y)$ and $K_\pi(y')$.
$\mathcal{C}_\gamma(y, y')$	Conditional executable mean successor cost, upper-bounding $\mathcal{W}(y, y')$ (Equation 9).
Δ_k	Realized pre-decision separation $d_Z(Z_k, Z'_k)$.
$\widetilde{d}_{Y,\xi}, \widetilde{D}_k$	Single paired-transition cost and its step-k rollout value: $\widetilde{D}_k = \alpha\Delta_k + (1-\alpha)\Delta_{k+1}$.
K, H, ω_k	Propagation-penalty window, evaluation/return horizon, and normalized penalty weight $1/K$.
R, R', R_{pair}	Branch task returns and the regularized paired reward.
$\lambda_{\mathrm{inv}}, \lambda_{\mathrm{ctr}}$	Initial-injection and affine-propagation penalty coefficients.
$\lambda_{\mathrm{abs}}, \lambda_A$	Propagation coefficients for the matched absolute and direct-state controls.
$\bar{q}, \bar{\rho}$	Training tolerance and target slope, selected on development data.
$V_k, \widehat{v}_k$	Positive affine violation $[\widetilde{D}_{k+1} - \bar{\rho}\widetilde{D}_k - \bar{q}]_+$ and its sample mean.
$\widehat{d}_0, n, \zeta$	Mean initial sample cost, independent held-out trajectory count, and failure probability for the confidence bound.
$\widehat{q}^{\mathrm{zero}}_{0.95}$	Empirical 95th-percentile post-decision zero-class leakage diagnostic.
$\widehat{b}_0, \widehat{b}_1$	Fitted mean intercept/slope for adjacent executable costs.
$\widehat{b}_{0,0.90}, \widehat{b}_{1,0.90}$	Fitted intercept/slope for the conditional 90th-percentile executable-cost profile.
$\widehat{a}_0, \widehat{g}_A$	Fitted intercept/gain for conditional successor-law OT on common pre-decision state pairs.
$q_A, \bar{q}_A$	Fixed direct-state diagnostic budget and direct-state training tolerance, respectively.
$\widehat{r}_{0.95}(q_A), \widehat{r}_{\max}(q_A)$	Quantile and observed maximum of $[\widehat{W}_i - q_A]_+/\Delta_i$ over the finite positive-pair panel.
$\widehat{q}^0_{A,0.95}$	Direct-state zero-pair leakage quantile, evaluated separately from positive-pair gains.
$u_+, \overline{u}_+, u_{+,0.90}$	Intercept-referenced exceedance $[\widetilde{D}_{k+1} - \widehat{b}_0]_+$ and its mean/90th percentile in the prediction features.
C, P, δ_C	Clean score, perturbed score, and allowed validation clean-score loss.

Context-dependent notation. A few symbols are overloaded, and their meaning is fixed by the context in which they appear. In the task tuple, P is the environment transition law and Q is the observation renderer; in evaluation, P denotes the perturbed score and Q may denote a common pair law. $T(o)$ denotes an observation transformation, whereas T in the decoding context denotes the sampling temperature. In rollout analysis γ denotes a coupling, while in return bounds it denotes the discount factor. The resolution threshold δ and the deployment-displacement bound δ are each defined within their own analyses and should be read only there. Finally, the coupling-gap bounds Δ_k of Proposition 6 are distinct from the realized state distances that share this symbol in the rollout notation.

C Operational Evaluation Protocol

This appendix fixes the data partitions, the intervention law, and the belief and history coordinates that every method shares; the transfer, mechanism, and selection results in the main text are evaluated under this common protocol.

C.1 Partitions and intervention sampling

Within each benchmark, training, development, and test task IDs are disjoint. ALFWorld draws 3,000 training instances and 256 further development instances from its training pool and reserves the 134 official unseen tasks for test. ScienceWorld uses 1,200 training variations, 256 development variations, and 300 held-out test variations stratified by task type. WebShop uses 1,000 training instructions, 256 development instructions, and 500 held-out test instructions, and all paraphrases of an instruction are confined to a single partition. For original τ-bench, all 165 published tasks are used for test only. Its training and development partitions are therefore study-specific: training uses 800 author-generated scenarios (400 per domain) and development uses 256 additional scenarios (128 per domain), all instantiated from the public schemas and policies without access to test user intents or expected action sequences. The generated pools are kept apart by separate database entity IDs and intent templates, whereas the shared public tools and policy documents define the common task family.

All methods receive the same partitions and the same demonstration pool. The 256 development instances are split into 128 for hyperparameter selection and 128 for calibration, and cross-fitting swaps these roles for auxiliary predictors. The checkpoint-selection ledger is computed on a separate validation panel of 256 draws taken with replacement from development tasks; repeated draws keep their original task as the independent cluster, and validation outcomes are excluded from inference on final test scores. The dedicated selection experiment scores every benchmark over a 24-action horizon, whereas the general task-score action limits are 50 for ALFWorld, 100 for ScienceWorld, 15 for WebShop, and 30 for original τ-bench. An invalid action consumes an action, and tool-return tokens consume context without consuming an extra action. Termination halts real simulator calls; absorbing padding enters only the trajectory statistics. Because the diagnostic horizon-24 WebShop episode extends beyond the 15-action primary score protocol, it is reported separately from primary P.

Eligible-opportunity mixture. For each task i, a frozen reference rollout supplies the pre-intervention prefix and every nonterminal checkpoint t within the first $\lfloor 0.6H_b \rfloor$ actions. At each such checkpoint, $\mathcal{V}_{it}$ denotes the finite set of valid atomic transformation instances that remain after the semantic and token-retention audits. Every eligible family contributes exactly four candidate instances from its frozen held-out generator, and duplicates are removed before counting. With $\mathcal{E}_i = \{t : |\mathcal{V}_{it}| > 0\}$ denoting its eligible checkpoints, the intervention law is

$$p_b(i, t, T) = w_{bi} \, \frac{\mathbf{1}\{t \in \mathcal{E}_i\}}{|\mathcal{E}_i|} \, \frac{\mathbf{1}\{T \in \mathcal{V}_{it}\}}{|\mathcal{V}_{it}|}. \tag{18}$$

For the first three benchmarks $w_{bi} = 1/N_b$; for τ-bench it is $1/(2N_{b,d(i)})$, which preserves equal domain weight. Replacing the last numerator by the number of valid instances from a given family yields the induced family weight. The total weight of a task is therefore fixed: neither a larger number of task opportunities nor additional paraphrase templates can increase it. A task whose $\mathcal{E}_i$ is empty receives the identity intervention and remains in the denominator, and the same rule covers early termination of the reference rollout. Eligible sets and prefix states are common to all methods, so the evaluated policy controls only the continuation after the selected checkpoint. Clean C uses the identical prefix together with an unmodified observation. C/P therefore evaluate reference-prefix continuations under the conventional benchmark score functions, whereas published end-to-end agents, which follow their own protocols, serve as contextual comparisons.

Task, checkpoint, and transformation draws are paired across methods and seeds. Primary P applies one atomic intervention; ordered compositions are evaluated separately as a stress test. The family-balanced diagnostic reverses the sampling order: it chooses the family uniformly first and then draws the task–checkpoint pair conditional on that family's eligibility. This conditioning is why family-balanced means need not reproduce primary P. Training applies the same sampling construction to training-only reference prefixes and generators. The full task reward covers both the shared prefix and the candidate-policy continuation.

C.2 Belief and retained-history coordinates

The coordinates defined here enter the block distances of the task-semantic metric in which all profiles are measured. Each weighted component listed in Appendix F is a block with a fixed predicate vocabulary. A state-only block uses its stated categorical, set, or scalar distance, whereas a block that includes hidden predicates or retained commitments uses

$$\delta_j = 0.50\, d_j^{\text{state}} + 0.25\, d_j^{\text{belief}} + 0.25\, d_j^{\text{history}}, \qquad d_j^{\text{belief}} = \frac{1}{|\mathcal{P}_j|} \sum_{p \in \mathcal{P}_j} |B_p(h, o) - B_p(h', o')|. \quad (19)$$

Because this decomposition operates inside each existing block weight, the outer weights still sum to one. The state-only blocks are inventory, irreversible failure/operation, and invalid-action affordances; every other block uses Equation 19. Empty predicate vocabularies are omitted when the schema is defined and are never created dynamically during scoring.

B_p is a frozen belief surrogate that conditions only on the retained canonical transcript and the current observation. Directly observed predicates receive probability zero or one. Unobserved predicates receive smoothed empirical frequencies computed from training simulator states consistent with the retained evidence, and an unseen evidence pattern falls back to the training task-type prior. Test latent values never update this surrogate. The resulting finite predicate representation serves as an approximate coordinate for the POMDP belief state. The history coordinate is the canonical set of still-active constraints, observed evidence records, and commitments, compared by Jaccard distance. A later explicit revocation removes the corresponding record, whereas an unmentioned constraint persists only while its supporting record remains in the retained context. The extractor reads only text retained after truncation.

The retained records are benchmark-specific: ALFWorld records discovered object locations and remaining goal predicates; ScienceWorld records observed measurements and apparatus prerequisites; WebShop records requested attributes, selected options, and cart commitments; and τ-bench records user-stated constraints, confirmations, and completed workflow steps. Metric validation is independent, using disjoint states and frozen extractors. As a consequence, two histories with the same latent state can have positive distance when their retained evidence or commitments differ, whereas paraphrases with identical extracted coordinates have distance zero. The learned-distance sensitivity experiment changes this measurement layer while retaining checkpoint access. The training regime evaluated here assumes restorable simulator and external-user random states together with an informative task-effect signal.

D Additional Theory and Estimation

Throughout this appendix, all state and action spaces are standard Borel, all kernels are measurable, and all bounded pseudometrics are Borel measurable. We further assume that the metric quotients admit measurable sections and that measurable optimal transport kernels exist for the costs used below. These conditions are satisfied by the finite spaces used in exact calibration; in continuous spaces, it suffices that the costs are bounded and continuous on compact metric quotients. Every disintegration and gluing step below is taken under these regularity conditions.

The appendix supplies the proofs behind the measurement theory (the affine profile and its perturbation recurrence, the environment–policy composition bound, and first-exit accounting) and behind the CRCO objective, and it establishes the theoretical objects on which the training and measurement protocols rely. The recurrence and estimation results come first and the coupling constructions follow, so that each implementation identity can be traced back to the assumptions it requires.

D.1 Proofs of the main bounds

D.2 Proofs connecting correspondence, the hinge, and task score

This subsection proves the results of the main text that underpin the CRCO objective: the correspondence-information proposition, which shows what the affine hinge measures beyond penalties on distance marginals, and the hinge-to-score theorem, which converts controlled hinge violations into a bound on task-score sensitivity.

Proof of Proposition 1. Set $a = \bar{\rho} d_\ell + \bar{q}$ and $b = \bar{\rho} d_h + \bar{q}$, so that $a \le b$. The key observation is that the difference of hinges $f(s) = [s - a]_+ - [s - b]_+$ is nondecreasing in its argument: it vanishes

below a, equals $s - a$ between a and b, and is constant at $b - a$ above b. Hence $f(s_h) \geq f(s_\ell)$, and rearranging this inequality yields Equation 11.

Strictness is witnessed by an explicit construction. Take $\bar{\rho} = 0.65$, $\bar{q} = 0.01$, $d_\ell = 0.10$, $d_h = 0.40$, $s_\ell = 0.05$, and $s_h = 0.20$. The aligned law places equal mass on $(d_\ell, s_\ell), (d_h, s_h)$, and the reassigned law places equal mass on $(d_\ell, s_h), (d_h, s_\ell)$. The two laws have identical current-distance and successor-distance marginals, so penalties that depend only on these marginals cannot separate them: both have absolute penalty 0.125, and for every fixed threshold t both have constant-hinge expectation $\frac{1}{2}([0.05 - t]_+ + [0.20 - t]_+)$. The affine hinge, by contrast, distinguishes them, with expectations 0 and $(0.20 - 0.075)/2 = 0.0625$, respectively. These values are exact properties of the mathematical construction, not empirical estimates. The construction concerns a positive-distance propagation step after injection and does not require a nonzero pre-intervention semantic distance. □

Proof of Theorem 2. By the definition of the positive part, every trajectory satisfies $\widetilde{D}_{j+1} \leq \bar{\rho}\widetilde{D}_j + \bar{q} + V_j$. Unrolling this inequality by repeated substitution gives $\widetilde{D}_H \leq B_H$, and because the metric is bounded, $\widetilde{D}_H \leq \min\{1, B_H\}$. Equation 4 then relates the endpoint distance to the rollout cost:

$$(1 - \alpha) d_Z(Z_{H+1}, Z'_{H+1}) \leq \widetilde{D}_H. \tag{20}$$

Because the coupling preserves both policy marginals, the clean and perturbed scores are expectations under the same coupled law, and therefore

$$\begin{aligned} |C_H - P_H| &\leq \mathbb{E}|g(Z_{H+1}) - g(Z'_{H+1})| \\ &\leq L_g \mathbb{E} d_Z(Z_{H+1}, Z'_{H+1}) \leq \frac{L_g}{1 - \alpha} \mathbb{E} \min\{1, B_H\}. \end{aligned} \tag{21}$$

The cap at one is supplied by the score range. For the finite-sample statement, fix (m, H): the n envelope values are IID and bounded in $[0, 1]$, so the one-sided Hoeffding inequality at error probability $\zeta/(MK)$, combined with a union bound, proves Equation 14. The union bound does not require independence across candidate policies and horizons, so the same tasks may be shared among them. Policy training and every choice of the metric, targets, and candidate set are made before this validation sample is drawn. □

Connection to the implemented uniform weights. In the theoretical envelope, a violation at step j carries weight $\bar{\rho}^{H-1-j}$, reflecting the number of remaining steps over which it can decay. The implemented reward instead retains uniform weights so that every scored edge is controlled. The uniform weighting still controls the envelope: for $H \leq K$,

$$\mathbb{E} B_H \leq \bar{\rho}^H \mathbb{E}\widetilde{D}_0 + \frac{\bar{q}(1 - \bar{\rho}^H)}{1 - \bar{\rho}} + K\,\mathbb{E}\left[\frac{1}{K}\sum_{j=0}^{K-1} V_j\right]. \tag{22}$$

Consequently, the two penalties in Equation 16 each control an explicit term of the score-sensitivity envelope, while the task reward prevents the policy from improving branch agreement by abandoning task completion. The statement is an implication from the objective to the outcome under the stated metric assumption; it does not claim that an optimizer reaches a global minimum.

Scope of the endpoint and selection statements. The index H counts post-decision costs: computing B_H requires $H + 1$ paired transitions, and the endpoint is Z_{H+1}. A five-step propagation probe therefore consumes six transitions. The bound extends to full-horizon terminal success only when the required violations are available through that endpoint or when an additional uniform continuation bound is established. For shorter probes, it certifies score sensitivity at the probe endpoint only, and full-horizon prediction is evaluated separately. Because the concentration event holds simultaneously, it remains valid when a policy is selected from the fixed M policies after validation. What it certifies, however, is the envelope estimates of these policies; it does not declare the selected policy test-optimal.

Corollary 1 (Persistent disturbances)**.** *Let $\mathcal{W}_k$ denote the Wasserstein distance between the two post-decision laws, and suppose that a disturbance kernel applied after step k adds at most η_k in Wasserstein distance. Then*

$$\mathcal{W}_k \leq \rho_\pi^k \mathcal{W}_0 + \sum_{j=0}^{k-1} \rho_\pi^{k-1-j}(q_\pi + \eta_j). \tag{23}$$

In particular, if $\eta_j \le \bar{\eta}$ at every step and $\rho_\pi < 1$, the asymptotic radius does not exceed $(q_\pi + \bar{\eta})/(1-\rho_\pi)$.

Proposition 2 (Return sensitivity)**.** *Suppose the per-step reward is L_r-Lipschitz under d_Y. Then, for remaining horizon H and discount γ,*

$$|J_H(\mu_0) - J_H(\nu_0)| \le L_r \sum_{j=0}^{H-1} \gamma^j \left(\rho_\pi^j \epsilon + q_\pi \sum_{i=0}^{j-1} \rho_\pi^i \right). \tag{24}$$

If, in addition, $\rho_\pi < 1$ and $\gamma < 1$, the right-hand side is at most $L_r\{\epsilon/[1-\gamma\rho_\pi] + q_\pi\gamma/[(1-\gamma)(1-\gamma\rho_\pi)]\}$.

The bound also covers binary terminal success: if states with different success labels are separated by a uniform distance margin $m_g > 0$, the success indicator has Lipschitz constant $1/m_g$. A Lipschitz goal-progress potential provides a dense surrogate for the same purpose.

Proof of Theorem 1. Fix any coupling of μ and ν. Applying Equation 8 conditionally to each paired state and integrating bounds the successor transport cost under the induced coupling, and taking the infimum over couplings yields $\mathcal{W}_{d_Y}(\mu K_\pi, \nu K_\pi) \le q_\pi + \rho_\pi \mathcal{W}_{d_Y}(\mu, \nu)$. Iterating this affine recurrence proves the theorem. □

Proof of Proposition 2. By Kantorovich duality, the expected reward difference at step j is at most L_r times the Wasserstein distance between the corresponding laws. Substituting the bound of Theorem 1 and summing over steps yields the finite-horizon expression, and summing the two resulting geometric series yields the stated infinite-horizon bound. □

Proof of Theorem 3. Assumption 2 imposes the following two uniform bounds on the components:

$$\mathcal{W}_{d_Z}(G(\cdot \mid y), G(\cdot \mid y')) \le q_G + L_G d_Y(y, y'), \tag{25}$$

$$\mathcal{W}_{d_Y}(S_\pi(\cdot \mid z), S_\pi(\cdot \mid z')) \le q_S + L_S d_Z(z, z'). \tag{26}$$

The proof also establishes the complementary pre-decision bound

$$\mathcal{W}_{d_Z}(H_\pi(z), H_\pi(z')) \le q_G + L_G q_S + L_G L_S d_Z(z, z'). \tag{27}$$

Fix y, y'. First couple $Z^+ \sim G(\cdot \mid y)$ and $Z'^+ \sim G(\cdot \mid y')$ optimally under d_Z; then, conditionally on (Z^+, Z'^+), couple the corresponding outputs of S_π optimally under d_Y. By the gluing lemma, the resulting joint law is a coupling of $GS_\pi(\cdot \mid y)$ and $GS_\pi(\cdot \mid y')$. Applying Equations 25–26 in turn gives

$$\mathcal{W}_{d_Y}(GS_\pi(y), GS_\pi(y')) \le q_S + L_S \mathbb{E} d_Z(Z^+, Z'^+) \tag{28}$$

$$\le q_S + L_S q_G + L_S L_G d_Y(y, y'). \tag{29}$$

Restricting to zero-distance pairs gives $q_\pi^0 \le q_S + L_S q_G$; with this intercept fixed, the supremum over positive-distance pairs yields the slope bound. Because the random next observation and the complete retained history are components of Z^+, the argument integrates over both. For the pre-decision bound, reverse the order: first couple $S_\pi(z), S_\pi(z')$ under d_Y, and then apply G conditionally. The resulting cost is at most $q_G + L_G \mathcal{W}_{d_Y}(S_\pi(z), S_\pi(z')) \le q_G + L_G q_S + L_G L_S d_Z(z, z')$, which proves Equation 27. This common upper envelope holds without requiring the empirical slopes measured in the two geometries to coincide. Finally, for the action refinement, couple actions under $c_{z,z'}$ and apply the pointwise d_Y bound together with the policy bound. This yields $q_S \le L_a q_{\text{policy}}$ and $L_S \le L_z + L_a L_{\text{policy}}$, and substituting these into the composition bound proves Equation 55. □

Proposition 3 (Aligned tightness of the typed product)**.** *Suppose that a positive-distance pair $(y_\star, y'_\star)$ admits an optimal G-coupling $\xi_\star$ attaining Equation 25, that Equation 26 is tight for $\xi_\star$-almost every successor pair, and that gluing the corresponding optimal S_π couplings yields an optimal coupling of $GS_\pi(y_\star), GS_\pi(y'_\star)$. Then the product slope bound of Theorem 3 is attained:*

$$\rho_\pi(q_S + L_S q_G) = L_G L_S. \tag{30}$$

Proof. Under these conditions the glued cost equals $q_S + L_S(q_G + L_G d_Y(y_\star, y'_\star))$, so the witness pair provides a lower bound that matches the upper bound supplied by Theorem 3. □

Theorem 4 (Exact and approximate quotient descent)**.** *Let $p : \mathcal{Y} \to \bar{\mathcal{Y}} = \mathcal{Y}/\!\sim$, where $y \sim y'$ iff $d_Y(y, y') = 0$, and assume that the standard-Borel quotient and its induced metric separate probability laws. Then $q_\pi^0 = 0$ iff $p_\# K_\pi(\cdot \mid y)$ does not depend on the representative y; equivalently, K_π induces a Markov kernel on $\bar{\mathcal{Y}}$. If instead $q_\pi^0 > 0$ and $\rho_\pi(q_\pi^0)$ is finite, every measurable section induces an approximate quotient kernel whose one-step ambiguity is at most q_π^0 and whose k-step ambiguity is at most*

$$q_\pi^0 \sum_{j=0}^{k-1} \rho_\pi(q_\pi^0)^j. \tag{31}$$

Proof of Theorem 4. Define the quotient metric by $\bar{d}([y], [y']) = d_Y(y, y')$. Pushforward to the quotient space preserves transport cost, $\mathcal{W}_{\bar{d}}(p_\#\mu, p_\#\nu) = \mathcal{W}_{d_Y}(\mu, \nu)$. Consequently, $q_\pi^0 = 0$ holds exactly when $p_\# K_\pi(\cdot \mid y) = p_\# K_\pi(\cdot \mid y')$ for every $y \sim y'$, which is precisely the condition under which

$$\bar{K}_\pi(\cdot \mid [y]) = p_\# K_\pi(\cdot \mid y) \tag{32}$$

does not depend on the representative and hence defines a Markov kernel on $\bar{\mathcal{Y}}$. The converse follows from the same identity.

Now suppose the leakage is nonzero. Fix a measurable section $s : \bar{\mathcal{Y}} \to \mathcal{Y}$ and define $\bar{K}_s(\cdot \mid \bar{y}) = p_\# K_\pi(\cdot \mid s(\bar{y}))$. Whenever $p(y) = \bar{y}$, we have $d_Y(y, s(\bar{y})) = 0$, and the definition of q_π^0 therefore gives $\mathcal{W}_{\bar{d}}(p_\# K_\pi(\cdot \mid y), \bar{K}_s(\cdot \mid \bar{y})) \le q_\pi^0$. Given any initial law μ_0, set $\mu_k = \mu_0 K_\pi^k$ and $\bar{\nu}_k = (p_\#\mu_0)\bar{K}_s^k$. To compare the two processes, lift an optimal quotient coupling by pairing each class on the second margin with its selected representative. Since representatives of the same class are at zero cost from each other, the lift does not increase the transport cost, and the affine profile implies

$$\mathcal{W}_{\bar{d}}(p_\#\mu_{k+1}, \bar{\nu}_{k+1}) \le q_\pi^0 + \rho_\pi(q_\pi^0)\mathcal{W}_{\bar{d}}(p_\#\mu_k, \bar{\nu}_k). \tag{33}$$

Because the initial discrepancy is zero, iterating this recurrence proves Equation 31. The bound is therefore uniform over the choice of section and initial law. □

Proposition 4 (Stability under equivalent semantic metrics)**.** *Let d_Y and d'_Y be pseudometrics on the same reachable set that are equivalent in the sense that $m d_Y \le d'_Y \le M d_Y$ for $0 < m \le M < \infty$. Then, for every $q \ge q^0_{\pi, d_Y}$,*

$$q^0_{\pi, d'_Y} \le M q^0_{\pi, d_Y}, \qquad \rho_{\pi, d'_Y}(Mq) \le \frac{M}{m}\rho_{\pi, d_Y}(q). \tag{34}$$

Moreover, the two metrics induce identical zero-distance equivalence classes.

Proof. For any laws μ, ν, transport costs are monotone in the ground cost, so $m\mathcal{W}_{d_Y}(\mu, \nu) \le \mathcal{W}_{d'_Y}(\mu, \nu) \le M\mathcal{W}_{d_Y}(\mu, \nu)$. Applying the upper inequality to zero-distance pairs proves the leakage bound. For every positive-distance pair, the two inequalities combine to give

$$\frac{[\mathcal{W}_{d'_Y}(K_\pi(y), K_\pi(y')) - Mq]_+}{d'_Y(y, y')} \le \frac{M}{m}\frac{[\mathcal{W}_{d_Y}(K_\pi(y), K_\pi(y')) - q]_+}{d_Y(y, y')}. \tag{35}$$

Taking the supremum over such pairs proves the amplification bound. □

Proposition 5 (Compositional semantic interventions)**.** *Consider a reachable chain of pre-decision states $z_0, \dots, z_m$ generated by successive interventions $z_i = T_i(z_{i-1})$, and define the induced policy-lift laws $\mu_i = S_\pi(\cdot \mid z_i)$ and their adjacent separations $\epsilon_i = \mathcal{W}_{d_Y}(\mu_{i-1}, \mu_i)$. Then*

$$\mathcal{W}_{d_Y}(\mu_0, \mu_m) \le \sum_{i=1}^{m} \epsilon_i. \tag{36}$$

If, in addition, K_π satisfies Assumption 1, then for every $k \ge 1$ the rollout separation obeys

$$\mathcal{W}_{d_Y}(\mu_0 K_\pi^k, \mu_m K_\pi^k) \le \rho_\pi^k \sum_{i=1}^{m} \epsilon_i + q_\pi \sum_{j=0}^{k-1} \rho_\pi^j. \tag{37}$$

The atomic sum can be sharpened for class-preserving interventions. Let $q_S^0 = \sup_{d_Z(z,z')=0} \mathcal{W}_{d_Y}(S_\pi(\cdot \mid z), S_\pi(\cdot \mid z'))$ *denote the zero-class leakage of the policy lift. If every* T_i *preserves the semantic class, transitivity yields the endpoint bound*

$$\mathcal{W}_{d_Y}(\mu_0, \mu_m) \le \min\left\{ q_S^0, \sum_{i=1}^{m} \epsilon_i \right\}. \tag{38}$$

Accordingly, the initial contribution in Equation 37 is at most $\rho_\pi^k \min\{q_S^0, \sum_i \epsilon_i\}$.

Proof. Equation 36 follows from the triangle inequality for the Wasserstein distance, applied along the intermediate policy-lift laws. By the measure-level form of Assumption 1, $\mathcal{W}_{d_Y}(\mu K_\pi, \nu K_\pi) \le q_\pi + \rho_\pi \mathcal{W}_{d_Y}(\mu, \nu)$. Applying this recurrence repeatedly to (μ_0, μ_m) and substituting Equation 36 proves Equation 37. If every adjacent distance is zero, the pseudometric triangle inequality further gives $d_Z(z_0, z_m) = 0$, so q_S^0 bounds the endpoint cost directly. The bounds serve different purposes: the atomic sum remains informative when only adjacent costs have been measured, whereas the class-preserving bound shows that semantic composition does not by itself force leakage to grow linearly with the number of transformations. □

Proposition 6 (Executable-to-Wasserstein transfer)**.** *Let* μ_k, ν_k *denote the post-decision marginal laws at step* k*, let* $\mathcal{W}_k = \mathcal{W}_{d_Y}(\mu_k, \nu_k)$ *be their theoretical Wasserstein separation, and let* $\mathcal{C}_{\gamma,k} = \mathbb{E}\widetilde{D}_k$ *be the executable mean cost under a feasible coupled trajectory. Suppose that the coupling gap satisfies* $0 \le \mathcal{C}_{\gamma,k} - \mathcal{W}_k \le \Delta_k$*, either exactly or on a simultaneous OT confidence event, and that the executable costs obey* $\mathcal{C}_{\gamma,k+1} \le \bar{q} + \bar{\rho}\mathcal{C}_{\gamma,k} + v_k$*. Then the theoretical separations satisfy*

$$\begin{aligned} \mathcal{W}_{k+1} &\le \bar{q} + \bar{\rho}\mathcal{W}_k + v_k + \bar{\rho}\Delta_k, \\ \mathcal{W}_H &\le \bar{\rho}^H \mathcal{W}_0 + \sum_{k=0}^{H-1} \bar{\rho}^{H-1-k}(\bar{q} + v_k + \bar{\rho}\Delta_k). \end{aligned} \tag{39}$$

Proof. Combine $\mathcal{W}_{k+1} \le \mathcal{C}_{\gamma,k+1}$, which holds by feasibility, with the gap bound $\mathcal{C}_{\gamma,k} \le \mathcal{W}_k + \Delta_k$, and iterate. The resulting statement is distributional: it holds under the audited pair law. The corresponding uniform quantity is obtained through the supremum over the reachable admissible set in Definition 1. □

Proposition 7 (Probe-to-occupancy transfer under covariate shift)**.** *Fix a measurable feasible outer coupling* $\gamma_{y,y'} \in \Gamma(K_\pi(\cdot \mid y), K_\pi(\cdot \mid y'))$*, and define its mean successor cost and affine residual by*

$$\mathcal{C}_\gamma^{\mathrm{out}}(y, y') = \mathbb{E}_{\gamma_{y,y'}} d_Y(Y^+, Y'^+), \qquad V_{\bar{q},\bar{\rho}}^{\mathrm{out}}(y, y') = \left[\mathcal{C}_\gamma^{\mathrm{out}}(y, y') - \bar{q} - \bar{\rho} d_Y(y, y')\right]_+. \tag{40}$$

Let P *be a probe law on reachable state pairs and* Q_0 *a target pair law, and define recursively* $Q_{k+1} = Q_k\gamma$*, the pair law obtained by executing the selected coupling. Suppose that for* $k = 0, \dots, H-1$ *the target laws are dominated by the probe law with bounded density ratio and that the probe residual is controlled:*

$$Q_k \ll P, \qquad \left\| \frac{dQ_k}{dP} \right\|_\infty \le C_k, \qquad \mathbb{E}_P V_{\bar{q},\bar{\rho}}^{\mathrm{out}} \le v. \tag{41}$$

Write $D_k = \mathbb{E}_{Q_k} d_Y(Y, Y')$ *and denote the two marginals of* Q_k *by* $Q_{k,Y}$ *and* $Q_{k,Y'}$*. Then the successor marginals and the selected outer pair process satisfy*

$$\mathbb{E}_{Q_k} \mathcal{W}_{d_Y}(K_\pi(\cdot \mid Y), K_\pi(\cdot \mid Y')) \le D_{k+1} \le \bar{q} + \bar{\rho} D_k + C_k v, \tag{42}$$

$$\mathcal{W}_{d_Y}(Q_{H,Y}, Q_{H,Y'}) \le D_H \le \bar{\rho}^H D_0 + \sum_{k=0}^{H-1} \bar{\rho}^{H-1-k}(\bar{q} + C_k v). \tag{43}$$

Moreover, if $V_{\bar{q},\bar{\rho}}^{\mathrm{out}} \in [0, 1]$ *and* $\widehat{v}$ *is the average over* n *independent* P*-probes fixed after policy training, then with probability at least* $1 - \zeta$ *the quantity* v *may be replaced by* $\widehat{v} + \sqrt{\log(1/\zeta)/(2n)}$.

Proof. Feasibility of the outer coupling gives $\mathcal{W}_{d_Y}(K_\pi(y), K_\pi(y')) \le \mathcal{C}_\gamma^{\mathrm{out}}(y, y')$, and the definition of the positive part gives $\mathcal{C}_\gamma^{\mathrm{out}} \le \bar{q} + \bar{\rho} d_Y + V_{\bar{q},\bar{\rho}}^{\mathrm{out}}$ pointwise. The density-ratio cap transfers the probe residual to the target law:

$$\mathbb{E}_{Q_k} V_{\bar{q},\bar{\rho}}^{\mathrm{out}} = \mathbb{E}_P\left[\frac{dQ_k}{dP} V_{\bar{q},\bar{\rho}}^{\mathrm{out}} \right] \le C_k \mathbb{E}_P V_{\bar{q},\bar{\rho}}^{\mathrm{out}} \le C_k v. \tag{44}$$

Taking expectations under Q_k proves Equation 42, where the identity $D_{k+1} = \mathbb{E}_{Q_k}\mathcal{C}_\gamma^{\mathrm{out}}$ follows from $Q_{k+1} = Q_k\gamma$. Iterating it proves the second inequality in Equation 43; the first holds because Q_H is itself a coupling of its two marginals. The finite-sample statement is the one-sided Hoeffding bound. The superscript marks this outer cost as distinct from $\mathcal{C}_\gamma$ in Equation 67, which additionally replaces the environment Wasserstein ground cost with an executable inner coupling. □

Proposition 7 makes explicit, through the density-ratio caps, how transfer from probe pairs to the occupancy law depends on support overlap. On enumerable kernels the caps C_k are available exactly; on real tasks, the evaluation instead reports cross-fitted density-ratio diagnostics and held-out coverage. This distinction matters for interpretation: the proposition requires a uniform ratio bound, whereas the empirical diagnostics summarize the sampled occupancy law.

Proposition 8 (Stopping-time completion of a local recurrence)**.** *Let $D_k \in [0, B]$ be an adapted divergence of a selected coupled pair process started in $\mathcal{R}$, and let $\tau_{\mathcal{R}} = \inf\{k \geq 1 : (Y_k, Y'_k) \notin \mathcal{R}\}$ be its first-exit time. The recoverable domain excludes terminal states and irreversible failures, and successful completion is absorbing. Fix deterministic $q, \rho \geq 0$ and suppose that the affine recurrence holds on survival:*

$$\mathbb{E}[D_{k+1}\mathbf{1}\{\tau_{\mathcal{R}} > k+1\} \mid \mathcal{F}_k] \leq \mathbf{1}\{\tau_{\mathcal{R}} > k\}(q + \rho D_k). \tag{45}$$

Then

$$\mathbb{E}D_H \leq \rho^H \mathbb{E}D_0 + q\sum_{j=0}^{H-1}\rho^j + B\Pr(\tau_{\mathcal{R}} \leq H). \tag{46}$$

A sharper completion accounts for the cause of exit. Let J label disjoint first-exit causes, and let deterministic $M_{t,j}$ upper-bound $\mathbb{E}[D_H \mid \tau_{\mathcal{R}} = t, J = j]$ on every positive-probability cell. Then

$$\mathbb{E}D_H \leq \rho^H \mathbb{E}D_0 + q\sum_{k=0}^{H-1}\rho^{H-1-k}\Pr(\tau_{\mathcal{R}} > k) + \sum_{t=1}^{H}\sum_{j}\Pr(\tau_{\mathcal{R}} = t, J = j)M_{t,j}. \tag{47}$$

The exit events are partitioned into successful pair completion and unresolved exit. If F_H denotes entry into an irreversible failure state by H, then no path with $\tau_{\mathcal{R}} > H$ or with a successful first exit has $F_H = 1$, and the failure probability decomposes over unresolved exits:

$$\Pr(F_H) = \sum_{t=1}^{H}\Pr(\tau_{\mathcal{R}} = t, J = \text{unresolved})\Pr(F_H \mid \tau_{\mathcal{R}} = t, J = \text{unresolved}). \tag{48}$$

Proof. Let $U_k = \mathbb{E}[D_k\mathbf{1}\{\tau_{\mathcal{R}} > k\}]$ denote the expected divergence on surviving paths. Taking expectations in the survival recurrence gives $U_{k+1} \leq q\Pr(\tau_{\mathcal{R}} > k) + \rho U_k$. Iterating this inequality and partitioning $\mathbb{E}D_H$ into surviving paths and the disjoint first-exit cells yields Equation 47. Bounding every survival probability by one and every $M_{t,j}$ by B recovers the coarser Equation 46. Applying the same partition to F_H proves Equation 48. □

Replacing the exit terms by statistical estimates yields calibrated predictors. These predictors become a probabilistic version of the bound when the stated upper-bound conditions hold jointly on a calibration event, and their empirical coverage is evaluated on independent complete trajectories.

Proposition 9 (Stability under deployment-kernel drift)**.** *Suppose a training-time kernel K satisfies $\mathcal{W}_{d_Y}(K(\cdot \mid y), K(\cdot \mid y')) \leq q + \rho d_Y(y, y')$ on the reachable set, and a deployed kernel K' is uniformly close to it in the sense that*

$$\delta = \sup_y \mathcal{W}_{d_Y}(K(\cdot \mid y), K'(\cdot \mid y)) < \infty, \tag{49}$$

then the deployed kernel inherits the profile with inflated leakage:

$$q^0_{K'} \leq q + 2\delta, \qquad \rho_{K'}(q + 2\delta) \leq \rho. \tag{50}$$

Consequently, whenever $\mathcal{W}_{d_Y}(\mu, \nu) \leq \epsilon$, the deployed rollout satisfies

$$\mathcal{W}_{d_Y}(\mu K'^k, \nu K'^k) \leq \rho^k\epsilon + (q + 2\delta)\sum_{j=0}^{k-1}\rho^j. \tag{51}$$

In the special case in which only the policy lift changes, with $\sup_z \mathcal{W}_{d_Y}(S_\pi(\cdot \mid z), S_{\pi'}(\cdot \mid z)) \leq \delta_S$, the kernels $K = GS_\pi$ and $K' = GS_{\pi'}$ satisfy Equation 49 with $\delta \leq \delta_S$.

Proof. Fix y, y'. By the triangle inequality for the Wasserstein distance,

$$\mathcal{W}_{d_Y}(K'(y), K'(y')) \le \mathcal{W}_{d_Y}(K'(y), K(y)) + \mathcal{W}_{d_Y}(K(y), K(y')) + \mathcal{W}_{d_Y}(K(y'), K'(y')) \quad (52)$$

$$\le q + 2\delta + \rho d_Y(y, y'). \quad (53)$$

Restricting first to zero-distance pairs and then to positive-distance pairs proves Equation 50, and iterating the measure-level affine inequality yields Equation 51. When only the policy lift changes, couple the two kernels through a shared $Z^+ \sim G(\cdot \mid y)$ and, conditionally on it, couple $S_\pi(\cdot \mid Z^+)$ with $S_{\pi'}(\cdot \mid Z^+)$. Integrating the resulting cost proves $\delta \le \delta_S$. □

Corollary 2 (Combined occupancy and deployment shift)**.** *Adopt the notation of Propositions 7 and 9. There exists a feasible deployed outer coupling $\gamma'_{y,y'}$ whose mean cost satisfies $\mathcal{C}^{\mathrm{out}}_{\gamma'}(y, y') \le \mathcal{C}^{\mathrm{out}}_{\gamma}(y, y') + 2\delta$. Suppose further that the deployed pair laws $Q'_k = Q'_0(\gamma')^k$ satisfy $Q'_k \ll P$ and $\|dQ'_k/dP\|_\infty \le C'_k$, and define $D'_k = \mathbb{E}_{Q'_k} d_Y(Y, Y')$. Then occupancy shift and deployment drift enter the bound additively:*

$$\mathcal{W}_{d_Y}(Q'_{H,Y}, Q'_{H,Y'}) \le D'_H \le \bar{\rho}^H D'_0 + \sum_{k=0}^{H-1} \bar{\rho}^{H-1-k}(\bar{q} + C'_k v + 2\delta). \quad (54)$$

Proof. Construct the deployed coupling by gluing an optimal $K'(y)$–$K(y)$ coupling, the selected $K(y)$–$K(y')$ coupling, and an optimal $K(y')$–$K'(y')$ coupling. By the triangle inequality, its expected endpoint cost is at most $\mathcal{C}^{\mathrm{out}}_{\gamma}(y, y') + 2\delta$. Repeating the density-ratio argument of Proposition 7 under Q'_k and iterating proves Equation 54. □

D.3 Action-effect refinement and native geometry

Proposition 10 (Native-metric compatibility)**.** *Fix G and $\alpha < 1$, and let d_Y^G denote the metric of Equation 3. Then Equation 25 holds with $q_G = 0$ and $L_G \le (1-\alpha)^{-1}$. This automatic bound, however, no longer identifies the comparative gain of G_v when the transition operator changes from G_0 to G_v while $d_Y^{G_0}$ is held fixed.*

The claim follows from the transition term of the native metric, which gives $\mathcal{W}_{d_Z}(G(y), G(y')) \le d_Y^G(y, y')/(1-\alpha)$, including on zero classes. The native geometry therefore guarantees compatibility with the environment kernel, whereas a frozen $d_Y^{G_0}$ is what makes environment interventions comparable (Appendix H). Equation 27 further places the common-pair diagnostic $\widehat{g}_A$ within the same typed analysis; this diagnostic uses the fixed d_Z geometry and involves no look-ahead overlap.

The coupling proof in Appendix D.1 integrates explicitly over the random next observation and the retained history. To make the action sensitivity inside L_S explicit, suppose that the action-effect metric and the policy satisfy $d_Y((z,a),(z',a')) \le L_z d_Z(z,z') + L_a c_{z,z'}(a,a')$ and $\mathcal{W}_c(\Pi_\pi(\cdot \mid z), \Pi_\pi(\cdot \mid z')) \le q_{\mathrm{policy}} + L_{\mathrm{policy}} d_Z(z,z')$, respectively. Then $q_S \le L_a q_{\mathrm{policy}}$ and $L_S \le L_z + L_a L_{\mathrm{policy}}$. Setting $q_\star = L_a q_{\mathrm{policy}} + (L_z + L_a L_{\mathrm{policy}}) q_G$, we obtain the explicit profile

$$q_\pi^0 \le q_\star, \qquad \rho_\pi(q_\star) \le L_G(L_z + L_a L_{\mathrm{policy}}). \quad (55)$$

The bound exposes two multiplicative interactions: inside the lift, action-effect sensitivity scales policy sensitivity, and the environment/history gain then multiplies the resulting discrepancy. This is the environment–policy composition bound of the measurement theory in refined form, and Section 5 evaluates these factors empirically under a fixed comparison geometry.

D.4 Tightness construction

Proposition 11 (General-kernel tightness and the unit-gain boundary)**.** *For every $q, \rho \ge 0$, every initial separation $\epsilon > 0$, and every finite horizon H, there exists a deterministic kernel on a pseudometric state space whose certifying profile is (q, ρ) and which attains Equation 10 with equality through H. With an aligned L_r-Lipschitz reward, the return bound of Equation 24 is attained as well. The construction is stated in arbitrary finite distance units; normalizing the metric to $[0,1]$ scales q and ϵ together and rescales the reward Lipschitz constant inversely.*

The constructions in this subsection show that the affine recurrence of the measurement theory is sharp within the stated settings. Each combines a semantic alias, to which the metric is blind, with an expanding coordinate. We first realize the sharp affine profile through the environment–policy composition GS_π under the native action-effect metric, and then give the general-kernel construction.

Proposition 12 (Realization in the native action-effect geometry)**.** *Fix $0 < \alpha < 1$, $q, \rho \geq 0$, and a finite horizon H. For every $\epsilon > 0$ satisfying $\epsilon \geq (1-\alpha)q$, there exist a deterministic agent and environment with binary actions that admit the native metric of Equation 3 and a certifying profile (q, ρ), together with a pair attaining Equation 10 through H. In this realization the minimum zero-class leakage is $(1-\alpha)q$, while q is the independently chosen affine budget. Distances are measured in arbitrary units, and normalization scales q and ϵ together.*

Proof. Let $z = (b, x)$, where $b \in \{0, 1\}$ is an observed representation alias and $x \in [-B, B]$, and set $d_Z(z, z') = |x - x'|$. Define

$$G((b,x),a) = \delta_{(b, f_a(x))}, \qquad f_a(x) = \mathrm{clip}_{[-B,B]}(\rho x + qa), \qquad \Pi_\pi(a \mid b, x) = \mathbf{1}\{a = b\}.$$

The alias is invisible to the semantic metric but determines the chosen action. Because both actions are allowed, the admissible post-decision set contains injected actions in addition to on-policy successors. Let $v = f_a(x) - f_{a'}(x')$ and $c = \alpha + (1-\alpha)\rho > 0$. The native distances then satisfy

$$d_Y(y, y') = \alpha|x - x'| + (1-\alpha)|v|, \quad d_Y(K_\pi y, K_\pi y') \leq c|v| + (1-\alpha)q.$$

Because $|v| \leq \rho|x - x'| + q$, subtracting gives $d_Y(K_\pi y, K_\pi y') - \rho d_Y(y, y') \leq q$ uniformly, clipping included. A zero-distance pair has $x = x'$ and $v = 0$; in the interior, a pair with opposite aliases and an identical current action attains successor distance $(1-\alpha)q$. Hence $q_\pi^0 = (1-\alpha)q \leq q$.

For the extremal pair, choose opposite aliases with their on-policy actions and aligned coordinate separation $\Delta_0 = (\epsilon - (1-\alpha)q)/c \geq 0$, and take B large enough that clipping is inactive through $H+1$. Then $\Delta_{k+1} = \rho\Delta_k + q$ and $D_k = c\Delta_k + (1-\alpha)q$, so that $D_0 = \epsilon$ and $D_{k+1} = \rho D_k + q$. Since positive-distance witnesses attain the uniform bound, $\rho_\pi(q) = \rho$. Finally, because the coordinate is confined to the bounded interval $[-B, B]$, the construction is compatible with the bounded-distance assumptions: dividing all distances by $2B$ places them in $[0, 1]$. □

The restriction on the initial distance is structural rather than an artifact of the proof: opposite on-policy aliases already differ in their impending action effects by $(1-\alpha)q$. The general-kernel construction below allows arbitrary initial separation precisely because it does not impose this look-ahead geometry. Both statements are sharp within their respective settings.

Proof of Proposition 11. Equip $\mathcal{Y} = \{0, 1\} \times [-B, B]$ with the pseudometric $d((b, x), (b', x')) = |x - x'|$, so that the first coordinate is a semantic alias invisible to the metric. Define the deterministic map $F(b, x) = (b, \mathrm{clip}_{[-B,B]}(\rho x + qb))$ and the kernel $K(\cdot \mid y) = \delta_{F(y)}$, and choose B large enough that clipping is inactive on the two witness trajectories through step H and on the interior neighborhood used below. Because scalar clipping is nonexpansive,

$$d(F(b,x), F(b',x')) \leq |\rho(x - x') + q(b - b')| \leq \rho d((b,x),(b',x')) + q. \tag{56}$$

The interior zero-distance states $(0, x)$ and $(1, x)$ are mapped exactly q apart, so $q^0 = q$; interior pairs with opposite aliases and aligned scalar difference attain the affine inequality, so $\rho(q) = \rho$. For the witness trajectories started at $(0, 0)$ and $(1, \epsilon)$, $D_{k+1} = q + \rho D_k$ as long as clipping remains inactive, and therefore $D_k = \rho^k \epsilon + q \sum_{j=0}^{k-1} \rho^j$ through step H. Finally, the reward $r(b, x) = L_r x$ on this finite reachable subset aligns every reward difference with D_k, so the finite-horizon return bound is attained. □

D.5 Finite-sample identification

Proposition 13 (Mean affine-summary concentration)**.** *Let $X_i = \widetilde{D}_i \in [0, 1]$ and $Z_i = \widetilde{D}_{i+1} = b_0 + b_1 X_i + e_i$ be n independent seed-level averages, each aggregating the same prespecified checkpoint and task schedule under a fixed executable coupling. Assume $\mathbb{E}[e_i \mid X_i] = 0$ and that the residuals are conditionally independent and σ-sub-Gaussian, and write $S_{xx} = \sum_i (X_i - \bar{X})^2 > 0$. Then, with probability at least $1 - \zeta$, the ordinary least-squares estimates $(\widehat{b}_0, \widehat{b}_1)$ satisfy*

$$|\widehat{b}_1 - b_1| \leq \sigma \sqrt{\frac{2\log(4/\zeta)}{S_{xx}}}, \tag{57}$$

$$|\widehat{b}_0 - b_0| \leq \sigma \sqrt{\frac{2\log(4/\zeta)}{n}} + \bar{X}\sigma \sqrt{\frac{2\log(4/\zeta)}{S_{xx}}}. \tag{58}$$

Proof. The slope error satisfies the identity $\widehat{b}_1 - b_1 = \sum_i (X_i - \bar{X})e_i / S_{xx}$, whose numerator is conditionally sub-Gaussian with variance proxy $\sigma^2 S_{xx}$; concentration then gives the slope bound. For the intercept, the identity $\widehat{b}_0 - b_0 = \bar{e} - (\widehat{b}_1 - b_1)\bar{X}$ combined with concentration of $\bar{e}$ and a union bound gives the intercept bound. □

Because the slope error scales with S_{xx}, it is the variation in current distances that determines slope precision. The probes sample equal numbers of checkpoints from prespecified $\widetilde{D}_0$ strata, which spreads the current distances across these strata. Confidence intervals on real tasks additionally account for the policy–task hierarchy.

D.6 One-sided witnesses against contraction

The upper bounds above have a counterpart in the opposite direction: any positive-distance pair whose successor Wasserstein distance is sufficiently large yields a lower bound on the global amplification coefficient. Independent OT estimates turn this observation into a finite-sample test.

Proposition 14 (Empirical-OT non-contraction witness)**.** *Fix a leakage budget $q \geq q_\pi^0$ independently of the positive-distance pair under test. Suppose $d = d_Y(y, y') > 0$ and that, on an event E, the empirical-OT estimate and the distance estimate satisfy $|\widehat{\mathcal{W}} - \mathcal{W}(y, y')| \leq e_W$ and $|\widehat{d} - d| \leq e_d$. Then, on E,*

$$\underline{\ell}_\rho(y, y'; q) = \frac{[\widehat{\mathcal{W}} - e_W - q]_+}{\widehat{d} + e_d} \leq \rho_\pi(q). \tag{59}$$

Consequently, $\underline{\ell}_\rho > 1$ refutes affine contraction at budget q. Analogously, for an exactly known zero-distance pair, $[\widehat{\mathcal{W}} - e_W]_+ \leq q_\pi^0$ provides a leakage witness.

Proof. On E, $[\widehat{\mathcal{W}} - e_W - q]_+ \leq [\mathcal{W}(y, y') - q]_+$ for the numerator and $d \leq \widehat{d} + e_d$ for the denominator. Dividing and invoking Equation 7 proves the positive-distance statement. For $d = 0$, the same lower confidence endpoint is at most $\mathcal{W}(y, y')$, which in turn is at most the supremum that defines q_π^0. □

The witness is valid only if candidate discovery is kept separate from confirmation. One split is used to rank pairs by a coarse local-gain score, after which at most 128 candidates are frozen; independent transition draws then compute $\widehat{\mathcal{W}}$, e_W, and e_d. Simultaneous one-sided radii allocate the error across the frozen candidates, and no confirmation sample is reused to choose a pair or tune q. The strength of the resulting statement depends on the setting. On enumerable kernels, exact reachability, exact distances, and finite-support concentration make Equation 59 a formal lower certificate. On real tasks, approximate reachability, inner OT, and a sampled pair pool reduce it to a one-sided observed-pair diagnostic: a confirmed violation is evidence against contraction on the audited distribution, whereas failing to find one is not evidence that the global supremum is below one.

D.7 A held-out confidence bound for the training surrogate

Proposition 15 (Hinge-to-trajectory confidence bound)**.** *Fix a policy and an executable coupling independently of a held-out evaluation sample, let $\widetilde{D}_k \in [0, 1]$, and define the positive affine violations $V_k = [\widetilde{D}_{k+1} - \bar{\rho}\widetilde{D}_k - \bar{q}]_+$ for $k = 0, \ldots, K-1$. Given n independent coupled trajectories, let $\widehat{d}_0 = n^{-1}\sum_i \widetilde{D}_{i0}$ denote the mean initial cost and $\widehat{v}_k = n^{-1}\sum_i V_{ik}$ the mean violations. Then, with probability at least $1 - \zeta$, simultaneously for every $H \leq K$,*

$$\mathbb{E}_{\gamma_{\mathrm{MC}}} \widetilde{D}_H \leq \bar{\rho}^H(\widehat{d}_0 + s_n) + \sum_{j=0}^{H-1} \bar{\rho}^{H-1-j}(\bar{q} + \widehat{v}_j + s_n), \qquad s_n = \sqrt{\frac{\log((K+1)/\zeta)}{2n}}. \tag{60}$$

Consequently, if the same violation bound holds uniformly at subsequent steps and $\bar{\rho} < 1$, then, apart from the decaying initial term, the sampled-coupling recurrence has asymptotic radius at most $(\bar{q} + \widehat{v} + s_n)/(1 - \bar{\rho})$.

Proof. Taking expectations in the pointwise inequality $\widetilde{D}_{k+1} \leq \bar{\rho}\widetilde{D}_k + \bar{q} + V_k$ yields the mean recurrence under the executable coupling. Since $\widetilde{D}_0, V_k \in [0, 1]$, one-sided Hoeffding bounds combined with a union bound over the $K + 1$ means provide the simultaneous events used in Equation 60. □

This bound converts the CRCO hinge into a held-out guarantee on mean divergence under the specified intervention law and executable coupling. Table 7 evaluates it for CRCO at $\zeta = 0.05$ on every benchmark, using IID task–simulator draws with replacement from the fixed held-out task pool of each seed policy; n therefore counts draws rather than distinct tasks. The diagnostic uses a common $(\bar{\rho}, \bar{q}) = (0.65, 0.01)$ across benchmarks and recomputes the violations at this tolerance, whereas the training tolerances remain benchmark-specific. At this common tolerance, the worst seed-specific upper bound remains below 0.33 on all benchmarks. These bounds lie well above the observed means in Table 7, so they should be read as conservative finite-sample guarantees rather than as estimates of the typical separation. Downstream score inference clusters repeated task draws. The remaining diagnostics address different questions: upper-tail profiles characterize variation within the evaluation law, while LOPO evaluates transfer across intervention families omitted from training.

Table 7: Five-step executable-coupling mean bounds for CRCO on each benchmark, evaluated at a common tolerance. n is the number of independent held-out task–simulator draws per fixed policy. Mean $\widetilde{D}_5$ is the observed separation averaged across six seeds; the upper bound is the worst seed-specific 95% bound from Equation 60, so every seed's bound lies at or below the reported value.

Benchmark	n per policy	Mean $\widetilde{D}_5$	95% upper bound
ALFWorld	500	0.045	0.244
ScienceWorld	300	0.061	0.321
WebShop	500	0.046	0.251
τ-bench	300	0.053	0.309

Proposition 16 (Executable-coupling CVaR recurrence)**.** *For $\alpha \in (0, 1)$, define the upper-tail risk measure (Rockafellar & Uryasev, 2000)*

$$\mathrm{CVaR}_\alpha(X) = \inf_{t \in \mathbb{R}} \left\{ t + \frac{1}{1-\alpha} \mathbb{E}[(X - t)_+] \right\}. \tag{61}$$

Fix a policy and an executable coupled trajectory, and define $V_k = [\widetilde{D}_{k+1} - \bar{\rho}\widetilde{D}_k - \bar{q}]_+$ for $k = 0, \ldots, H-1$. Then the pathwise hinge identity implies

$$\mathrm{CVaR}_\alpha(\widetilde{D}_H) \le \bar{\rho}^H \, \mathrm{CVaR}_\alpha(\widetilde{D}_0) + \sum_{k=0}^{H-1} \bar{\rho}^{H-1-k} \left(\bar{q} + \mathrm{CVaR}_\alpha(V_k) \right). \tag{62}$$

If, moreover, $\mathrm{CVaR}_\alpha(V_k) \le v_\alpha$ uniformly and $\bar{\rho} < 1$, then once the initial term has decayed, the tail radius is at most $(\bar{q} + v_\alpha)/(1 - \bar{\rho})$.

Proof. The definition of V_k gives, on every coupled trajectory, $\widetilde{D}_{k+1} \le \bar{q} + \bar{\rho}\widetilde{D}_k + V_k$. Because upper-tail CVaR is monotone, translation equivariant, subadditive, and positively homogeneous, it follows that $\mathrm{CVaR}_\alpha(\widetilde{D}_{k+1}) \le \bar{q} + \bar{\rho}\,\mathrm{CVaR}_\alpha(\widetilde{D}_k) + \mathrm{CVaR}_\alpha(V_k)$, and iterating this inequality proves Equation 62. □

Whereas the preceding bound controls mean divergence, this result controls upper-tail risk under the specified executable coupling and perturbation law. Empirically, we estimate $\mathrm{CVaR}_{0.95}$ from held-out trajectories within each policy via the Rockafellar–Uryasev empirical objective, treating the training seed as the outer unit of uncertainty. Table 30 evaluates the recurrence under infrequent high-gain disturbances.

D.8 Predictable state-dependent gains and deterministic envelopes

The affine profile summarizes amplification by a single slope, whereas local gains may depend on the state. For $\widetilde{D}_k > \delta$, define the empirical residual local gain

$$r_k = \frac{[\widetilde{D}_{k+1} - \widehat{b}_0]_+}{\widetilde{D}_k}. \tag{63}$$

The reporting threshold $\delta = 10^{-3}$ is fixed before evaluation. Now let the coefficients q_k, ρ_k be nonnegative and $\mathcal{F}_k$-measurable. If only the conditional-mean condition $\mathbb{E}[\widetilde{D}_{k+1} \mid \mathcal{F}_k] \le q_k + \rho_k \widetilde{D}_k$

is assumed, the type-correct conclusion is the one-step recursion $\mathbb{E}\widetilde{D}_{k+1} \le \mathbb{E}q_k + \mathbb{E}[\rho_k \widetilde{D}_k]$, which implies neither independence between coefficients and distances nor any factorization of their product. When deterministic envelopes satisfy $q_k \le \bar{q}_k$ and $\rho_k \le \bar{\rho}_k$ almost surely, iteration yields

$$\mathbb{E}\widetilde{D}_H \le \mathbb{E}\widetilde{D}_0 \prod_{j=0}^{H-1} \bar{\rho}_j + \sum_{i=0}^{H-1} \bar{q}_i \prod_{j=i+1}^{H-1} \bar{\rho}_j. \tag{64}$$

Under the stronger pathwise condition that $\widetilde{D}_{k+1} \le q_k + \rho_k \widetilde{D}_k$ almost surely, the corresponding random-coefficient statement is instead

$$\mathbb{E}\widetilde{D}_H \le \mathbb{E}\left[\widetilde{D}_0 \prod_{j=0}^{H-1} \rho_j + \sum_{i=0}^{H-1} q_i \prod_{j=i+1}^{H-1} \rho_j \right], \tag{65}$$

where the products remain inside the expectation. Equations 64 and 65 follow by monotonicity and by pathwise substitution, respectively. Because neither statement reduces the random gains to a single fitted coefficient, the empirical analysis reports the full distribution of r_k alongside the fitted affine slope.

D.9 Sample costs and the two coupling layers

The Wasserstein term in Equation 3 is a theoretical quantity that is not directly observable. Given an executable coupling $\xi_G(dz^+, dz'^+ \mid y, y')$ of the two environment kernels, one paired transition provides the look-ahead sample cost

$$\widetilde{d}_{Y,\xi}(y, y'; z^+, z'^+) = \alpha d_Z(z, z') + (1-\alpha) d_Z(z^+, z'^+), \tag{66}$$

which upper-bounds the theoretical cost in expectation, $d_Y(y, y') \le \mathbb{E}_{\xi_G} \widetilde{d}_{Y,\xi}$, with equality requiring an optimal environment coupling. Training uses this observable sample cost, while the Wasserstein term itself is estimated offline by empirical OT.

The rollout cost adds an outer coupling layer to this environment coupling. For a current pair (y, y'), let γ_K couple the successor post-decision states through the policy and environment sampler. Given $(U, U') \sim \gamma_K$, let $(V, V') \sim \xi_G(\cdot \mid U, U')$ denote the coupled look-ahead pre-decision states, and write $U = (U_Z, U_A)$ and $U' = (U'_Z, U'_A)$. The theoretical Wasserstein response and the executable mean cost are then

$$\begin{aligned} &\mathcal{W}(y, y') = \mathcal{W}_{d_Y}(K_\pi(\cdot \mid y), K_\pi(\cdot \mid y')), \\ &\mathcal{C}_\gamma(y, y') = \mathbb{E}_{\substack{(U,U') \sim \gamma_K \\ (V,V') \sim \xi_G(\cdot \mid U, U')}} [\alpha d_Z(U_Z, U'_Z) + (1-\alpha) d_Z(V, V')] = \mathbb{E}[\widetilde{D}_{k+1} \mid Y_k = y, Y'_k = y'], \end{aligned} \tag{67}$$

where $\gamma = (\gamma_K, \xi_G)$ collects both executable couplings. The corresponding two-layer upper bound is Equation 9, and equality throughout requires optimal couplings at both layers. Accordingly, the quantities are measured differently: the observed $\widetilde{D}_{k+1}$ has conditional mean $\mathcal{C}_\gamma(y, y')$, whereas $\mathcal{W}$ is estimated offline by empirical optimal transport over cloned action effects.

D.10 Executable coupling bound

Let $\Gamma(\mu, \nu)$ denote the set of all couplings of the successor distributions, so that by definition $\mathcal{W}_d(\mu, \nu) = \inf_{\gamma \in \Gamma(\mu,\nu)} \mathbb{E}_\gamma d(Y, Y')$. The rollout sampler of CRCO is one feasible element of this set: tokenwise maximal coupling followed by common-random simulator execution defines γ_{MC}. Its paired look-ahead sample therefore has expectation $\mathcal{C}_{\gamma_{\mathrm{MC}}} \ge \mathcal{W}$, with slack at an outer and an inner level: the outer inequality arises because a merely feasible coupling of successor laws is used, and the inner one because the environment Wasserstein distance in Equation 3 is replaced by a feasible transition coupling. When token laws are identical, maximal coupling emits identical sequences almost surely and thereby avoids the positive cost that a product coupling would incur. Training penalizes violations measured under this executable coupling, and Proposition 6 transfers the resulting recurrence to the theoretical Wasserstein profile while explicitly accounting for the stepwise feasible-coupling gaps. These gaps are estimated by diagnostic empirical OT from additional cloned executions, whose cost is charged to evaluation compute.

D.11 JOINT MAXIMAL-COUPLING LIKELIHOOD

Fix one token position and condition on the two generated prefixes. Let p_v and p'_v denote the corresponding policy probabilities, $m_v = \min(p_v, p'_v)$ their overlap, and $M = \sum_v m_v$ the total overlap mass. The tokenwise maximal-coupling probability mass function is then

$$Q_\theta(v, w) = \begin{cases} m_v, & v = w, \\ \dfrac{(p_v - m_v)(p'_w - m_w)}{1 - M}, & v \neq w \text{ and } M < 1, \\ 0, & \text{otherwise}, \end{cases} \tag{68}$$

with $Q_\theta(v, v) = p_v$ when $M = 1$. Because the residual supports are disjoint, the off-diagonal branch cannot emit equal tokens. Treating EOS as absorbing, the probability of an action pair is $q_\theta(a, a') = \prod_\ell Q_\theta(v_\ell, v'_\ell \mid v_{<\ell}, v'_{<\ell})$. This distribution has marginals $\pi_\theta(a \mid z)$ and $\pi_\theta(a' \mid z')$, and it coalesces identical conditional laws almost surely.

Let $\mathcal{T}_i$ denote the complete paired continuation from a frozen intervention checkpoint through the task-return horizon. The coupling is maintained at every action, not only within the regularized prefix, and the first $K + 1$ paired transitions supply $\widetilde{D}_0{:}\widetilde{D}_K$. Conditional on the frozen checkpoint, the continuation's policy likelihood factor and the resulting trajectory ratio are

$$L_\theta(\mathcal{T}_i) = \prod_{t=0}^{H_i - 1} q_\theta(a_{it}, a'_{it} \mid z_{it}, z'_{it}), \qquad r_i = \frac{L_\theta(\mathcal{T}_i)}{L_{\text{old}}(\mathcal{T}_i)}. \tag{69}$$

Absorbing null actions contribute a factor of one. The paired GRPO surrogate combines this trajectory ratio with the pair-level advantage A_i:

$$\mathcal{L}_{\text{joint}} = -\mathbb{E}_i[\min(r_i A_i, \text{clip}(r_i, 1 - \epsilon_{\text{clip}}, 1 + \epsilon_{\text{clip}}) A_i)] + \frac{\beta_{\text{KL}}}{2} \mathbb{E}_i[\text{KL}_i + \text{KL}'_i]. \tag{70}$$

Each KL term averages the marginal tokenwise divergence to the frozen reference over its continuation. The environment-coupling factors cancel from the ratio because they do not depend on θ. Ratios are computed from sums of joint log probabilities over all sampled actions, including the actions after the regularized prefix, since these also affect task return. Automatic differentiation follows Equation 68; at incidental ties $p_v = p'_v$ we use the midpoint subgradient, and for identical conditioning inputs the two derivative paths coincide.

A change in the sign of $p_v - p'_v$ can change the residual support of Q_θ, so Equation 70 is not used as a global importance-sampling identity. It is instead defined as a PPO-style local surrogate over samples drawn from Q_{old} and is refreshed rather than reused: we run one minibatch epoch and refresh the coupled rollouts after every update. To check that updates remain local, a vocabulary-wide audit computes the new-coupling mass outside the old support, the fraction of sampled token pairs assigned zero new mass, and the usual clip fraction. For Qwen and Gemma, respectively, the median changed-support mass is 0.004 and 0.005, with 95th percentiles of 0.018 and 0.021; sampled pairs with zero new mass account for 0.03% and 0.04%; and the clip fractions are 8.7% and 9.4%. Together, these diagnostics establish that training operates in the intended local update regime.

E COMPLETE TRAINING ALGORITHM

The listing below states one CRCO iteration in execution order. A task and step are drawn from the frozen training reference-prefix pool, and the replayed checkpoint is cloned into a clean and a transformed branch. Both branches then advance under maximal coupling of their next-token laws and common simulator randomness, which yields the per-step successor distances; terminated branches are absorbed and padded to the fixed horizon. Only after both branches reach the task-return horizon are pair returns computed and standardized within each four-pair group, and these advantages drive the clipped policy update. The shuffled-threshold control used in the state-correspondence experiments differs only in the update step.

Algorithm 1 Complete Coupled-Rollout Contraction Optimization (CRCO) training loop. Each iteration branches a checkpointed training prefix into clean and transformed observations, advances both branches under maximal coupling with common simulator randomness while recording the successor distances, standardizes pair returns within each four-pair task–checkpoint group, and applies a clipped GRPO update against a frozen reference.

Input: policy π_θ; frozen reference π_{ref}; checkpointable environments; $\mathcal{T}_{\mathrm{train}}$; depth K; targets $(\bar{\rho}, \bar{q})$; weights $(\lambda_{\mathrm{inv}}, \lambda_{\mathrm{ctr}})$; $\omega_k = 1/K$
Sample task and step τ from the frozen training reference-prefix pool.
Replay and checkpoint (x_τ, h_τ); clone it into branches o_τ and $T(o_\tau)$.
for $k = 0, \ldots, K$ **do**
    Maximally couple the two next-token laws at temperature 0.7 and top-$p = 1$.
    Execute the paired actions with common simulator randomness and record $\widetilde{D}_k$.
    if either branch has terminated **then**
        Map it to its absorbing terminal state and append null actions to the fixed horizon.
    end if
end for
Continue both branches to the task-return horizon; invalid actions follow benchmark dynamics.
for each four-pair task–checkpoint group **do**
    Compute the pair return in Equation 16.
    Form $A_i = (R_i - \overline{R}_{-i})/(s_R + 10^{-6})$ over the four coupled pairs.
end for
Optimize Equation 70 with the ratio in Equation 69, $\epsilon_{\mathrm{clip}} = 0.2$, one minibatch epoch, gradient clipping 1.0, and $\beta_{\mathrm{KL}} = 0.02$.
Refresh the coupled rollout batch; keep the pre-CRCO GRPO reference frozen.
Shuffled-threshold control: detach donor thresholds and update only recipient ratios.

F Metric Construction and Validation

This appendix specifies and validates the policy-independent task-semantic geometry on which the measurement theory and every reported profile rest. The measurement protocol fixes this geometry before any policy comparison. Equation 2 defines the complete real-task distance, and the weights below are its outer block weights; Appendix C.2 expands the belief and retained-history blocks into their exact subweights and extraction rules. Frozen parsers map $z = (x, h, o)$ to goal propositions, belief-relevant evidence, retained constraints and commitments, task-state coordinates, and executable action effects. Set-valued coordinates use the Jaccard distance $|A \triangle B|/|A \cup B|$, with empty–empty distance zero; categorical coordinates use the discrete metric, and ordered progress uses the normalized absolute difference. Belief coordinates compare the Bernoulli distributions of each fixed task predicate under the Wasserstein distance on $\{0, 1\}$, averaged over predicates. Parsing removes surface differences in token order, whitespace, field order, tool aliases, and transcript wording, and histories have $d_Z = 0$ when all extracted task-relevant coordinates agree. The latent distance d_X is the renormalized restriction to coordinates computable from x alone.

The expanded d_Z coordinate weights are fixed per benchmark and sum to one in each case. ALFWorld assigns 0.40 to goal predicates, 0.25 to object relations, 0.15 to inventory, 0.10 to valid affordances, and 0.10 to irreversible failures. ScienceWorld assigns 0.35, 0.25, 0.15, 0.10, 0.05, and 0.10 to goal predicates, apparatus state, inventory, relations, affordances, and failures, respectively. WebShop assigns 0.35, 0.20, 0.15, 0.10, 0.10, and 0.10 to attribute satisfaction, selected-product identity, cart commitments, page/filter state, valid actions, and purchase error. τ-bench assigns 0.30, 0.30, 0.15, 0.10, 0.10, and 0.05 to user constraints, database delta, workflow commitments, valid tools, policy compliance, and irreversible operation. During metric construction a coordinate may be read from the latent simulator, but belief and history affect d_Z only when they change one of these canonical task-relevant quantities. All components lie in $[0, 1]$, and $\alpha = 0.30$ in Equation 3.

We validate the metric against absolute differences in continuation value. Continuation values come from exact dynamic programming or the benchmark oracle where available and otherwise from 64 Monte Carlo continuations per state under a fixed reference policy; weights are selected on training states only. The resulting Spearman correlations are 0.82 ± 0.02 on ALFWorld, 0.76 ± 0.03 on ScienceWorld, 0.71 ± 0.03 on WebShop, and 0.68 ± 0.04 on τ-bench, where uncertainty is a task-bootstrap standard deviation conditional on the frozen reference policy.

The post-decision metric satisfies symmetry and the triangle inequality by construction. Two sensitivity tests then probe its weighting and its reliance on latent predicates. Random Dirichlet perturbations around the fixed component weights keep P within 52.7–53.6. Separately trained observation-metric policies (Table 3) reach P of 53.0/55.3 and $\widehat{g}_A = 0.74/0.78$ on the Qwen/Gemma anchors, slightly below the predicate metric's 53.3/55.7 and 0.71/0.74. The conclusion is therefore tied neither to one favorable weighting nor to latent predicate access.

F.1 INDEPENDENT METRIC TRIANGULATION

To test sensitivity to the feature representation itself, not only to its weights, we evaluate four independently constructed distances, each normalized on training tasks and frozen before evaluation. The oracle-value distance is the absolute difference between continuation values estimated without paired-policy outcomes. The learned bisimulation distance is trained only on training tasks. For the human distance, three annotators blinded to method, action string, and downstream outcome compare 600 individual states against fixed task-effect anchors; an ordinal mixed-effects model estimates latent severities $h(y)$ and defines $d_{\text{human}}(y, y') = |h(y) - h(y')|$, which is a pseudometric. Character-level edit distance on serialized traces serves as a surface-form negative control. All profile fits, failure models, and CRCO–GRPO contrasts are rerun without retuning policies or metric-specific thresholds. Uncertainty resamples trained policies as the outer unit, and the human-distance audit additionally resamples annotators and anchors.

Table 8: Metric triangulation, macro-averaged across four benchmarks. Under every task-semantic distance, CRCO lowers the slope relative to paired GRPO by a similar margin and failure AUROC stays close to that of the primary metric, whereas the text-edit negative control agrees poorly with value and human severity and shows little slope reduction. Value and human columns are Spearman correlations; rank agreement compares checkpoint slope rankings with those of the primary metric. The slope pair reports paired-GRPO/CRCO on the common eight-seed panel, and reduction is one minus their ratio.

Distance	Value-gap agreement	Human-severity agreement	Slope-rank agreement	Failure AUROC	Slope pair	Reduction
Primary state-effect metric	0.74	0.79	1.00	**0.82**	1.13/0.69	**38.9%**
Oracle continuation value	1.00	0.76	0.91	0.80	1.14/0.72	36.8%
Training-split bisimulation	0.77	0.75	0.88	0.81	1.09/0.71	34.9%
Blinded human effect severity	0.73	0.86	0.84	0.79	1.04/0.69	33.7%
Text edit distance (negative control)	0.19	0.17	0.31	0.62	1.07/0.98	8.4%

The four task-semantic distances yield consistent checkpoint rankings, held-out failure AUROC of 0.79–0.82, and 33.7–38.9% lower slopes for CRCO; each reduction is directly reproducible from the displayed slope pair. Text edit distance gives substantially weaker rankings and prediction. Because the semantic distances are constructed independently, their agreement supports the robustness of the empirical result to the metric representation, even though their zero-distance classes may differ.

Observation-trained optimization without privileged predicates. The primary implementation measures training-time action effects with simulator predicates. The observation-trained ablation removes this privileged input: it fits a deep bisimulation encoder (Zhang et al., 2021) on disjoint training-task tuples (o_t, a_t, r_t, o_{t+1}), freezes it before policy optimization, and computes every paired rollout cost in that latent pseudometric. Its intervention pool contains only renderer operations whose invariance follows from the interface definition—unordered-key permutations, serialization whitespace, and registered tool aliases—so no latent-state, oracle-belief, or per-action next-state audit enters either training or test admission. Table 3 reports separately optimized policies rather than rescored predicate-trained checkpoints. The retained 2.6-point mean P gain over matched absolute propagation shows that the optimization effect survives the removal of privileged metric inputs. Blinded-human distances offer a lower-throughput measurement alternative, and verifier-derived state deltas can be used when executable success constraints are available. Every variant, however, still requires checkpointable or replayable training environments to form paired continuations; the paper makes no claim for non-replayable training. Deployment always uses standard observations and a single rollout.

The metric-triangulation table, by contrast, is a separate representation audit: it re-evaluates frozen checkpoints under four independently constructed distances, again without retuning policies or metric-specific thresholds. The resulting real-task profiles characterize propagation under the chosen metric and probe distribution. They are diagnostics for that audited distribution, not estimates of a global Wasserstein supremum or of a model-independent contraction constant.

F.2 ACTION-DISTRIBUTION AUDIT

This audit is used for diagnostics only. Injection OT samples executable actions from each branch, evaluates their cloned effects, merges duplicate canonical effects, and solves the outer transport problem. Closed-loop-profile OT instead draws outer samples $U_i \sim K_\pi(\cdot \mid y)$ and $U'_j \sim K_\pi(\cdot \mid y')$; its ground matrix uses the exact $d_Y(U_i, U'_j)$ in deterministic benchmarks, whereas in stochastic kernels each entry is an inner 32-sample OT estimate of the transition term in Equation 3. Outer sample counts are 8, 16, 32, and 64, and a sensitivity check raises the inner samples to 64. A same-observation test compares maximal coupling, independent product coupling, entropy-matched product coupling, greedy decoding, and empirical OT, and transformed-observation probes repeat the comparison.

Table 9: Fixed-policy coupling audit on deterministic simulator states. On identical inputs, maximal coupling and greedy decoding incur zero cost at the smallest sample budget, whereas both product couplings retain positive cost and yield higher mean slopes. All rows except the greedy control sample actions at $T = 0.7$. Each call is one successor-law comparison; profiles additionally require their current-state costs.

Estimator	Samples / calls	Same-observation cost	Mean slope
Maximal coupling, $T = 0.7$	1 / 2	0.000	0.72
Independent product	1 / 2	0.087	0.78
Entropy-matched product	1 / 2	0.083	0.77
Greedy paired control	1 / 2	0.000	0.69
Empirical OT	16 / 32	0.009	0.73
Empirical OT	32 / 64	0.006	0.71
Empirical OT	64 / 128	0.005	0.70

The product-coupling cost remains positive even after action entropy is matched, whereas maximal coupling and greedy decoding give zero cost on identical inputs. Increasing empirical OT from 32 to 64 samples changes the slope by approximately 0.01. These diagnostic calls are charged separately from training and deployment. Positive finite-sample OT can arise between independent samples from identical stochastic laws even when the population distance is zero.

The normalized coupling-gap diagnostic compares the mean executable injection cost with empirical action-effect OT at $T = 0.7$ on transformed-observation probes: $(\widehat{C}^{\mathrm{inj}}_\gamma - \widehat{W}^{\mathrm{inj}}_{\mathrm{OT}})/(\widehat{W}^{\mathrm{inj}}_{\mathrm{OT}} + 0.05)$, with both terms computed under the same input-pair law. The empirical difference estimates a feasible-coupling gap, and sampling error need not preserve its population nonnegativity. Sinkhorn-8 draws eight effect samples in total (four per marginal), solves an entropically regularized transport plan following the Sinkhorn formulation (Cuturi, 2013), and samples one pair for the rollout, with regularization selected on disjoint calibration states. It is a training sampler that is better aligned with action effects but more expensive.

Table 10: Training-coupling frontier of P, slope, and coupling gap against simulator cost. Among samplers at unit cost, tokenwise maximal coupling gives the highest P and the smallest gap; the effect-aware Sinkhorn sampler improves all three further at four times the cost. P and mean slope use greedy deployment of each trained variant, whereas the gap uses its stochastic $T = 0.7$ injection laws. The simulator multiplier is relative to one two-branch transition.

Coupling	P ↑	Mean slope ↓	Norm. gap ↓	Simulator calls
Independent product	51.0	0.74	0.073	1.0×
Entropy-matched product	51.6	0.73	0.068	1.0×
Tokenwise maximal	53.3	0.69	0.021	1.0×
Effect-aware Sinkhorn-8	**54.0**	0.66	0.010	4.0×
Empirical OT-64 (diagnostic)	–	–	0.000	64.0×

Relative to tokenwise maximal coupling, Sinkhorn-8 gains a further 0.7 P points and reduces the coupling gap, but at four times the simulator cost. Maximal coupling therefore sets the primary method's trade-off between transport accuracy and training cost: it removes most of the gap left by product couplings without additional simulator calls.

F.3 TRAINING-TO-DEPLOYMENT KERNEL AUDIT

Policy optimization samples both marginals at $T = 0.7$, whereas the primary C/P results and real-task profiles use the frozen greedy deployment policy; this audit checks whether the change of kernel affects the conclusions. On disjoint probes across the eight confirmatory seeds and saved checkpoints, the sampling and greedy kernels give $(\widehat{q}^{\mathrm{zero}}_{0.95}, \widehat{b}_0, \widehat{b}_1) = (0.014, 0.021, 0.72)/(0.012, 0.018, 0.69)$ and upper-quantile slopes of $0.86/0.83$, respectively. Cross-kernel displacement is $\widehat{\delta}_{0.95} = 0.018$, checkpoint-rank agreement is 0.94, and the method ordering is unchanged.

F.4 CONDITIONAL STATE-DYNAMICS ESTIMATION

Write $\Delta_k = d_Z(Z_k, Z'_k)$, reserving A_k for the sampled action. By Equation 4, $\widetilde{D}_k = \alpha\Delta_k + (1-\alpha)\Delta_{k+1}$, so adjacent regression pairs share one step-distance. In full generality,

$$\begin{aligned}\mathrm{Cov}(\widetilde{D}_k, \widetilde{D}_{k+1}) = \alpha^2 \mathrm{Cov}(\Delta_k, \Delta_{k+1}) + \alpha(1-\alpha)\mathrm{Cov}(\Delta_k, \Delta_{k+2}) \\ + \alpha(1-\alpha)\mathrm{Var}(\Delta_{k+1}) + (1-\alpha)^2 \mathrm{Cov}(\Delta_{k+1}, \Delta_{k+2}). \end{aligned} \tag{71}$$

Consequently, even for IID Δ_k with positive variance, the population OLS slope is nonzero: $\alpha(1-\alpha)/\{\alpha^2 + (1-\alpha)^2\} = 0.362$ at $\alpha = 0.3$. This value is an analytic null for the overlapping statistic, whereas the dynamics themselves are identified by direct transition-law measurements.

For Equation 74, a frozen reference-policy mixture supplies 384 predecision pairs per benchmark, stratified by distance decile, remaining horizon, recoverability, and irreversible-risk predicate, and all methods use these same pairs. At stochastic states, 32 independent environment/user draws per marginal under the greedy policy form the empirical-OT ground matrix, and a 64-draw sensitivity analysis uses fresh samples; deterministic states use their exact successors. Nonnegative mean and 0.90-quantile regressions map the current Δ to successor-law OT. Intervals resample training seeds and common pairs jointly across methods. Outcome tasks are disjoint, and prediction is clustered by fitted policy profile.

Table 11: Independent-dynamics audit. Among the four training methods, CRCO attains the lowest overlapping slope and fixed-pair gain and the highest LOMO and LOBO AUROC, while the two nulls recover the analytic overlap value and chance-level prediction, respectively. $\widehat{b}_1$ uses adjacent executable costs, whereas $\widehat{g}_A$ uses fixed-pair conditional OT. AUROC uses the direct-state profile; the two null rows test different estimators.

Method	Overlap. $\widehat{b}_1$	Fixed-pair $\widehat{g}_A$	LOMO AUROC	LOBO AUROC	Rank agreement
Paired GRPO	1.13	1.08	0.748	0.701	0.91
One-step consistency	1.10	1.04	0.773	0.736	0.90
Matched absolute propagation	0.81	0.86	0.812	0.758	0.92
CRCO	**0.69**	**0.71**	**0.845**	**0.804**	0.95
IID scalar-distance null	0.361	–	–	–	–
Conditional-law shuffle	–	0.004	0.51	0.50	0.03

Each method contributes $4 \times 8 \times 8 = 256$ benchmark–seed–checkpoint units. The final column compares, within method, the checkpoint rankings induced by the overlapping slope and by the direct-state gain; pooled over methods, rank agreement is 0.93. The CRCO–matched-absolute contrast in $\widehat{g}_A$ is -0.15 $[-0.21, -0.09]$, a 17.4% relative reduction, and doubling the stochastic successor samples changes it by only 0.006. Each row's AUROCs evaluate its tasks with predictions trained without its method family (LOMO) or its benchmark (LOBO).

The IID null independently resamples scalar distances from each policy's pooled marginal before constructing the overlapping costs; by deliberately removing temporal dependence, it recovers 0.361, in line with the analytic null above. Separately, the conditional-law null randomly reassigns complete successor-cloud pairs across policies and current pairs within each benchmark, without distance stratification. This breaks the association with current distance and method while retaining the benchmark's response distribution, and $\widehat{g}_A$ falls to 0.004 $[0.000, 0.015]$. Null AUROCs use only the shuffled profile, with benchmark offsets removed inside each training fold; shared capability controls are excluded from this null because they retain predictive information. These estimator nulls complement the fixed-input policy comparison.

Direct propagation and training without look-ahead mixing. This analysis tests whether the benefit of state-conditioned propagation depends on the look-ahead mixing in the rollout-sample distance. On the same pair panel, we fit $\mathbb{E}[\Delta^+ \mid z, z']$ under common-random-number (CRN) execution and then replace that executable response by conditional OT. We also train a direct-state variant with

$$R^{\mathrm{direct}}_{\mathrm{pair}} = \frac{R + R'}{2} - \lambda_{\mathrm{inv}} \widetilde{D}_0 - \frac{\lambda_A}{K} \sum_{k=1}^{K} [\Delta_{k+1} - \bar{\rho}\Delta_k - \bar{q}_A]_+. \tag{72}$$

This variant uses the same $Z_0{:}Z_{K+1}$ rollouts, injection term, task reward, eight seeds, and coefficient-search budget. Skipping the initial zero-distance pair keeps the propagation penalty separate from the injection term. The direct-state tolerance $\bar{q}_A$ is selected under the same development-only budget as $\bar{q}$. Only the propagation functional changes, and evaluation still uses the frozen d_Z.

Table 12: Direct-state replication on common pairs against the RQ2 controls. Both state-conditioned objectives, the direct-state affine hinge and CRCO, exceed the matched absolute, constant-threshold, and shuffled-current controls in C and P and attain lower CRN and OT slopes. CRN and OT slopes fit $\Delta \to \Delta^+$ with identical inputs; training the direct-state hinge tests whether the gains depend on look-ahead mixing.

Training objective	C	P	CRN slope	OT $\widehat{g}_A$
Matched absolute propagation	60.0	50.7	0.92	0.86
Constant-threshold hinge	60.4	51.2	0.86	0.83
Shuffled-current hinge	60.2	50.9	0.91	0.85
Direct-state affine hinge	61.3	53.1	**0.77**	0.73
CRCO	**61.6**	**53.3**	0.78	**0.71**

Trained without look-ahead mixing, the direct-state variant gains 2.4 P points $[1.1, 3.7]$ over matched absolute propagation, retaining 92.3% of the full 2.6-point margin, and differs from CRCO by -0.2 $[-0.8, 0.4]$. Across objectives, CRN slopes exceed conditional-OT slopes by 0.03–0.07, as expected for a feasible rather than optimized coupling. The broad ordering is preserved, with one small reversal between the direct-state affine hinge and CRCO: CRN favors the former by 0.01, whereas conditional OT favors CRCO by 0.02. Direct-state training thus retains the advantage. Together with the activation controls, these results support state-conditioned propagation as the useful algorithmic structure across distance constructions.

G Profile Estimators and Validation Accounting

This appendix specifies the measurements behind the main-text profile, transfer, and selection figures: the estimator definitions, the common-budget comparison of propagation profiles, the paired cross-backbone transfer intervals, and the simulator-call ledgers. The two checkpoint-selection studies keep separate candidate pools and separate resource accounts.

G.1 Distributional profile estimators

Real-task probes estimate the profiles under specified reachable pair distributions, and the estimators mirror the profile's separation of quotient leakage from residual amplification. Dedicated pairs with $d_Y \le \delta$ estimate the zero-distance leakage quantile $q^{\mathrm{zero}}_{0.95}$, whereas positive-distance probes estimate the nonnegative mean affine summary

$$\mathbb{E}[\widetilde{D}_{k+1} \mid \widetilde{D}_k] \approx b_0 + b_1 \widetilde{D}_k, \qquad b_0, b_1 \ge 0, \tag{73}$$

together with a separately fitted conditional 90th-quantile profile $(b_{0,0.90}, b_{1,0.90})$. Ratios are computed only on distances above the prespecified resolution δ. Hats denote fitted distributional summaries, whereas (q_π, ρ_π) denotes the uniform theoretical profile.

Because adjacent $\widetilde{D}_k$ share a look-ahead state, the pre-decision kernel and a non-overlapping profile are defined jointly:

$$H_\pi(dz^+ \mid z) = \int \Pi_\pi(da \mid z) G(dz^+ \mid z, a), \quad \mathcal{W}_{d_Z}(H_\pi(z), H_\pi(z')) \approx a_0 + g_A d_Z(z, z'), \tag{74}$$

The gain $\widehat{g}_A$ is estimated by conditional empirical OT from independent successor draws on a frozen pair law Q shared by all methods. For each input pair, the response is computed from fresh transition

Table 13: Common-budget direct-state propagation on input pairs shared by all methods. CRCO has the lowest point estimate in every profile column, from intercept and mean gain to the zero-distance tail and the residual gains, while held-out coverage is similar across methods. Fits are performed within benchmark and seed before averaging; residual gains use $q_A = 0.04$, and maxima refer to the finite panel.

Method	$\widehat{a}_0$	$\widehat{g}_A$	$\widehat{q}^0_{A,0.95}$	$\widehat{r}_{0.95}$	$\widehat{r}_{\max}$	Held-out coverage
Paired GRPO	0.031	1.08	0.049	1.44	1.83	93.7%
Matched absolute	0.022	0.86	0.021	1.17	1.49	94.1%
Constant hinge	0.024	0.83	0.020	1.12	1.45	94.5%
CRCO	0.019	0.71	0.018	0.95	1.23	94.3%

draws, so no realized successor is reused as the predictor, and fixing Q standardizes occupancy across methods. The resulting fixed-pair dynamic gain is a diagnostic of H_π, whose uniform bound is given in Theorem 3.

G.2 Common-budget affine envelopes

The common-budget comparison uses the same 384 positive-distance pairs per benchmark as direct-state OT. For each pair, write $\Delta_i = d_Z(z_i, z_i')$ and let $\widehat{W}_i$ denote empirical OT between $H_\pi(z_i)$ and $H_\pi(z_i')$. Each input pair receives fresh successor samples, and all methods share the input law. Nonnegative least squares fits $(\widehat{a}_0, \widehat{g}_A)$. Separately, the budget $q_A \in \{0.02, 0.04, 0.06\}$ is fixed before any candidate policy is evaluated, and the residual gains are

$$R_i(q_A) = \frac{[\widehat{W}_i - q_A]_+}{\Delta_i}, \qquad \widehat{r}_{0.95}(q_A) = \mathrm{Quantile}_{0.95}\{R_i(q_A)\}, \qquad \widehat{r}_{\max}(q_A) = \max_i R_i(q_A), \tag{75}$$

computed over pairs with $\Delta_i \geq 0.05$, with resolution checks at 0.025 and 0.10. These budgets are diagnostic constants in the normalized d_Z geometry and are distinct from the training $\bar{q}$ and from the post-decision zero-class leakage quantile. Zero-distance pre-decision probes are reported separately as $\widehat{q}^0_{A,0.95}$; a budget below their tail still yields a valid positive-pair diagnostic but provides no zero-class coverage.

Figure 3A plots the mean and tail gains, and Table 13 gives the complete numerical profile.

Coverage is the proportion of 384 additional independent input-pair draws that satisfy $\widehat{W} \leq q_A + \widehat{r}_{0.95}\Delta$, computed with the same finite-sample OT estimator used for fitting and validation; it therefore measures held-out empirical-OT response coverage. Uncertainty in method contrasts is obtained by resampling the eight seed blocks and then the paired input panels, which keeps the training instance as the outer independent unit. Against matched absolute propagation, the CRCO-minus-absolute gain difference is -0.15 with interval $[-0.21, -0.09]$, and the fixed-budget tail-gain difference is -0.22 with interval $[-0.31, -0.11]$; both intervals exclude zero and establish the profile improvement. The intercept difference is -0.003 with interval $[-0.007, 0.001]$: its point estimate is smaller, but its interval includes zero.

The method ranking is the same at every budget. At budgets 0.02/0.04/0.06, the tail gains are 1.30/1.17/1.08 for matched absolute, 1.27/1.12/1.02 for constant hinge, and 1.09/0.95/0.86 for CRCO; the corresponding maximum gains are 1.71/1.49/1.35, 1.66/1.45/1.31, and 1.47/1.23/1.10. Monotonicity within each row is required by Equation 75, so the experimental comparison is the method ranking across budgets, in which CRCO is lowest and constant hinge second throughout. Observed maxima above one persist even where the 95th-percentile envelope lies below one, because the finite panel retains rare expanding pairs. Long-horizon prediction therefore combines these profiles with the separately calibrated competing-exit terms and retains all paths after first exit.

G.3 Cross-backbone paired contrasts

The focused transfer comparison uses six independently initialized runs per method and benchmark, and its matched objectives share the same six seed blocks. Within each comparison, backbone-specific initialization, token cap, optimizer FLOPs, injection coefficient, coupling, and checkpoint rule are held constant, and regularization strength is tuned on the same nested validation grid. Under these controls, the CRCO-minus-absolute P intervals are $[3.2, 7.2]$ for Qwen2.5-14B and $[0.7, 4.0]$ for Llama-3.1-8B, and the corresponding intervals against the constant hinge are $[2.2, 6.0]$ and $[0.4, 3.3]$, respectively; every interval excludes zero, although the Llama margins are smaller. The corresponding

Gemma-3-12B-IT contrast is reported in the aligned main matrix. Paired inference treats seed as the outer unit, nests tasks within benchmark, and analyzes each backbone separately. Table 1 contains the complete six-objective rows; this section supplies the transfer intervals and accounting details.

G.4 Checkpoint-selection accounting

Let C_{clean} denote the shared clean-validation simulator cost per seed. Including this cost, the checkpoint-selection saving of profile recurrence over full-horizon validation, whose per-panel logical-slot budgets appear in Table 18, is

$$\frac{49{,}152 - 28{,}672}{49{,}152 + C_{\text{clean}}}. \tag{76}$$

With a matched 24-action clean ledger, $C_{\text{clean}} = 49{,}152$ logical slots, which gives a 20.8% total validation-slot saving; the 41.7% figure is the fixed-horizon logical-slot reduction before the clean ledger is counted. In the primary eight-seed CRCO selection experiment, the instrumented ledgers record 43,250 full-horizon calls, 26,680 profile calls, and 44,010 shared clean calls. These counts correspond to a 38.3% incremental robustness-selection saving and a 19.0% saving after the clean ledger is added. The separate 28-panel unshuffled direct-state pool keeps its own call ledger in Table 18. Simulator-call and training-cost accounting are reported independently: on the common profiling stack, normalized training cost is $1.05\times$ that of paired GRPO and $1.91\times$ that of the unpaired reference.

H Component Interventions and Composition Tests

This appendix tests the component structure of the measurement theory under controlled conditions. Orthogonal interventions first vary action effects or environment feedback while the task geometry is held fixed; finite-kernel suites then test when the composed gain attains its product bound (aligned kernels) and how much slack it leaves otherwise (non-aligned kernels).

Orthogonal component interventions. The action-effect experiment alters only the consequence multiplier of one designated nonoptimal action; across the resulting $L_a = 0.42, 0.66, 0.89$, the CRCO gain grows from 3.8 to 11.8 to 17.2 points. The environment/history intervention first freezes a reference comparison geometry: d_Z^{ref} and every action-effect ground cost are constructed from the native task semantics, after which d_Y^{ref} is defined once through the native kernel in Equation 3. The intervention then alters only the transition operator G_v, producing a reversible wrapper that partially restores the prior state, the unchanged native kernel, and an amplified wrapper that repeats the state-dependent consequence once before rendering. Actions, rewards, rendering length, $\widetilde{D}_0$ strata, and the frozen metrics are held fixed. Because $\alpha < 1$, $d_Y^{\text{ref}}(y, y') = 0$ implies zero native successor distance; because both wrappers are deterministic functions of the same canonical state and action effect, they also map every zero class identically. Hence $q_G = 0$ for all three variants, and the manipulated quantity is residual environment gain measured against a common task geometry.

Table 14: Controlled environment/history-gain intervention with quotient leakage held at zero. Rows match initial distance and action-effect distribution, and Gain is CRCO P minus GRPO P. The gain increases with residual environment gain across the three variants under both geometries. Primary L_G is computed in one frozen reference geometry; the endogenous geometry, which rebuilds the metric within each variant, is a sensitivity analysis.

Environment variant	L_G^{ref}	L_G^{endog}	GRPO P	CRCO P	Gain
Restoring wrapper	0.62	0.59	53.6	57.9	+4.3
Native dynamics	0.82	0.82	41.5	53.3	+11.8
Amplified feedback	1.04	1.09	29.7	46.1	+16.4

The fixed-geometry and endogenous estimates induce identical rank orders (Spearman $r_s = 1.00$), so the ordering of the variants does not depend on the choice of geometry. Equal-weight regressions of the three displayed cell-mean gains yield $+28.7$ P points per unit L_G^{ref} (policy-clustered 95% interval $[22.4, 35.0]$) and $+24.0$ per unit L_G^{endog} ($[18.5, 29.6]$); neither interval contains zero. A prespecified policy-level regression containing L_G^{ref}, L_a, method, and both method interactions estimates adjusted component sensitivities, whereas the displayed coefficients retain the directly reproducible cell-mean

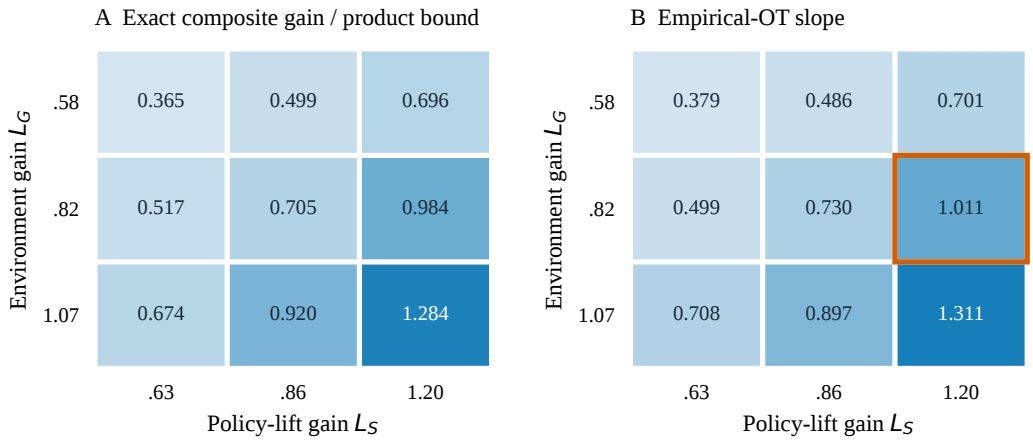


Figure 6: Aligned-witness composition across environment and policy-lift gains. **A:** Exact composite gain, which equals the theorem's product bound by construction. **B:** Empirical-OT slope on the same cells, which reproduces the exact pattern except at the unit boundary. Both panels share the 0–1.4 color scale and print every value; the outlined cell is the one that crosses the unit boundary under estimation (0.984 exact versus 1.011 estimated).

estimand. The finite-kernel experiments below test both the product inequality and its equality conditions.

The finite-kernel comparison consists of an aligned and a non-aligned suite. The aligned suite crosses three environment gains with three policy-lift gains at quotient leakage 0.01, and every kernel satisfies the common-witness condition in Equation 30: one reachable direction attains both component bounds and the composed optimal transport cost. Each cell contains 40 independently generated kernels, evaluated on 512 stratified state pairs. Equality then predicts unit exponents in the log-linear fit $\log \widehat{\rho} = \beta_0 + \beta_G \log L_G + \beta_S \log L_S$, namely $\beta_G = \beta_S = 1$.

Across the aligned cells, exact enumeration gives $\rho_K/(L_G L_S) = 1.000$, as the common-witness construction requires. Empirical OT tracks these exact values with MAE 0.021, Spearman $r_s = 0.983$, and unit-boundary accuracy $8/9$; its only boundary error is the 0.984 cell adjacent to the boundary. The fitted exponents, $\widehat{\beta}_G = 1.01$ (95% interval $[0.96, 1.07]$) and $\widehat{\beta}_S = 1.00$ ($[0.94, 1.07]$) with $R^2 = 0.994$, both have intervals that contain the predicted unit value. Recomputing leave-cell-out prediction from the nine displayed cell means, the log-linear product model attains original-scale RMSE 0.028, compared with 0.091 for an ordinary least-squares additive model with intercept and predictors (L_G, L_S); the multiplicative form therefore fits the aligned cells better than an additive composition.

The non-aligned suite is generated independently and imposes no common witness: it rotates or permutes gain directions, places the component suprema on disjoint reachable pairs, or mixes state-dependent modes. Exact enumeration computes ρ_K and $L_G L_S$ separately before empirical OT is run, and the resulting tightness ratio quantifies the slack that a generic composition leaves relative to its componentwise worst-case bound.

Table 15: Product-bound audit on independently generated non-aligned finite kernels. Tightness is exact $\rho_K/(L_G L_S)$, so values below one measure slack; violations are counted by exact enumeration before empirical estimation. No suite violates the bound, and every suite leaves substantial median slack.

Suite	Kernels	Median	IQR	90th pct.	OT MAE	Violations
Rotated/permuted directions	40	0.78	[0.68,0.87]	0.94	0.026	0/40
Disjoint witness pairs	40	0.62	[0.53,0.73]	0.82	0.029	0/40
State-dependent mixtures	40	0.72	[0.61,0.83]	0.92	0.030	0/40
All non-aligned kernels	120	0.72	[0.61,0.84]	0.93	0.028	0/120

The upper bound is never violated, yet the median non-aligned composition attains only 72% of it. The experiment therefore establishes both sides of the typed law: equality is recoverable under the stated alignment conditions, and strict slack is typical when the components' worst-case directions differ.

I STATE CORRESPONDENCE, RECOVERY, AND SELECTION

Figure 5 summarizes the recovery and decision outcomes; this appendix supplies the complete interventions, numerical records, and candidate-pool accounting behind it. The analysis isolates two ingredients of the CRCO mechanism: correspondence between the current state pair and its propagation threshold, tested by threshold reassignment, and the opportunity to recover before an irreversible exit, tested by varying repair availability. Checkpoints from the resulting unshuffled arm then support a separate evaluation of selection decisions.

I.1 GRADED INTERVENTION ON STATE CORRESPONDENCE

The mechanism hypothesis is that matching the affine threshold to the current separation identifies the steps at which recovery is possible. The competing explanation is that any threshold distribution with comparable activation and mean penalty supplies the same selective regularization. We test this on the direct-state affine objective, so that the manipulated threshold and the independently measured response use the same state distance without look-ahead overlap.

Table 16: Graded state-correspondence intervention on the direct-state affine objective. Reassign is fixed by the transposition/derangement design; C/P are clean/perturbed scores; $\widetilde{D}_0$ measures initial injection; $\widehat{g}_A$ is the fixed-pair direct-state gain; and the final columns give complete-cohort successful first exits and irreversible failures. In every backbone–task cell, the unshuffled arm attains the highest P and successful-exit rate and the lowest gain and failure rate, while the half-reassignment response is heterogeneous across cells.

Objective	Reassign	C / P	$\widetilde{D}_0$	$\widehat{g}_A$	Successful exit	Failure
Qwen2.5-7B / ScienceWorld						
Matched absolute	–	46.4 / 37.5	.119	.90	39.8%	15.1%
Direct affine, unshuffled	0%	47.6 / 41.1	.117	.74	47.6%	10.3%
Direct affine, half	50%	47.8 / 38.9	.116	.90	42.7%	13.8%
Direct affine, full	100%	47.1 / 37.9	.119	.91	41.6%	14.2%
Qwen2.5-7B / WebShop						
Matched absolute	–	71.1 / 61.0	.112	.82	56.0%	8.8%
Direct affine, unshuffled	0%	71.9 / 63.0	.111	.69	58.7%	6.9%
Direct affine, half	50%	71.5 / 62.5	.109	.73	58.1%	7.4%
Direct affine, full	100%	71.3 / 61.3	.112	.79	56.7%	8.4%
Gemma-3-12B / ScienceWorld						
Matched absolute	–	48.4 / 39.7	.111	.96	41.7%	14.4%
Direct affine, unshuffled	0%	49.1 / 43.4	.109	.72	49.1%	9.2%
Direct affine, half	50%	48.6 / 40.8	.107	.91	42.5%	12.7%
Direct affine, full	100%	48.8 / 41.0	.110	.88	43.2%	12.8%
Gemma-3-12B / WebShop						
Matched absolute	–	72.8 / 63.3	.106	.88	57.0%	8.4%
Direct affine, unshuffled	0%	73.4 / 65.4	.104	.71	60.2%	6.5%
Direct affine, half	50%	73.6 / 64.6	.105	.78	59.0%	7.2%
Direct affine, full	100%	72.9 / 64.3	.103	.82	58.5%	7.7%

Each update group contains four trajectory pairs. The half-reassignment arm applies one uniformly selected transposition, changing exactly two of the four assignments, whereas the full arm samples a derangement, changing all four. Both operations preserve the current-distance multiset, the successor distances, the paired branches, and the task rewards, so the manipulation targets the correspondence between each pair and its threshold. Donor thresholds are detached old-rollout labels, and each pair updates only its recipient likelihood; the intervention is therefore a recipient-only surrogate rather than the gradient of a permutation-coupled group return. The design crosses Qwen2.5-7B-Instruct and Gemma-3-12B-IT with ScienceWorld and WebShop, and matched absolute propagation anchors task performance. Qwen uses eight shared initialization blocks and Gemma six; every arm shares the same reference-prefix pool, injection coefficient, update schedule, and rollout budget.

Development calibration approximately matches regularization strength: on held-out batches, the affine arms span 38.2–40.1% activation and mean penalty 0.026–0.030. Initial $\widetilde{D}_0$ varies by at most 0.003 within a backbone–task cell and shows no common directional change, so the arms differ

in threshold correspondence rather than in initial injection. Relative to the unshuffled direct-state affine objective, full reassignment reduces P by 2.8 points on ScienceWorld (paired interval $[1.6, 4.0]$ across backbone–seed blocks) and by 1.4 points on WebShop ($[0.5, 2.3]$), while the corresponding $\widehat{g}_A$ increases by 0.17 and 0.11. Recovery degrades accordingly: successful exits fall by 6.0 points on ScienceWorld and 1.9 points on WebShop, and irreversible failure rises by 3.8 and 1.4 points. The half arm lies between the aligned and fully reassigned arms in three cells but slightly overshoots the full arm on Gemma ScienceWorld, so the finite-sample response is heterogeneous rather than mechanically proportional to the reassignment fraction. Preserving the threshold multiset under closely matched held-out strength therefore does not preserve propagation or recovery: current–successor correspondence carries independent information.

I.2 Recovery opportunity as a moderator

This intervention varies the availability of valid repair transitions while calibrating baseline return within each irreversible-exit stratum. The enumerable-kernel experiment crosses recovery probability $r \in \{0.15, 0.45, 0.75\}$ with irreversible-exit risk $e \in \{0.04, 0.16\}$ and adjusts neutral transition mass on development kernels. In ScienceWorld, a frozen wrapper changes repair-transition availability and inserts a reward-neutral detour so that matched-absolute P remains comparable within each exit regime; the high-exit wrapper additionally redirects a fixed fraction of failed repair attempts to an absorbing failure state. The reported estimand is this baseline-matched joint wrapper, including its finite-horizon changes in path length and reachability. Wrapper parameters are frozen before affine checkpoints are scored, and metrics, action sets, initial-distance strata, and training budgets remain fixed.

Table 17: Recovery-opportunity interventions. Outcomes are terminal return on a 0–100 display in panel A and ScienceWorld P in panel B. Δ is the direct-state affine hinge minus matched absolute propagation; gain and failure columns report absolute/affine values. In both panels the affine advantage is positive in every cell and largest when repair transitions are plentiful and irreversible-exit risk is low.

Exit condition	Recovery opportunity	Absolute outcome	Affine outcome	Δ	Gain (abs./aff.)	Failure (abs./aff.)
A. Enumerable kernels with initial injection fixed at 0.020						
$e = 0.04$	$r = 0.15$	61.2	64.0	+2.8	.86 / .78	13.1% / 11.6%
$e = 0.04$	$r = 0.45$	60.5	66.9	+6.4	.89 / .69	12.8% / 9.7%
$e = 0.04$	$r = 0.75$	61.1	68.7	+7.6	.85 / .65	13.4% / 8.9%
$e = 0.16$	$r = 0.15$	52.7	54.7	+2.0	.90 / .81	27.5% / 26.4%
$e = 0.16$	$r = 0.45$	52.2	56.8	+4.6	.93 / .73	27.1% / 23.9%
$e = 0.16$	$r = 0.75$	52.9	56.9	+4.0	.87 / .72	28.2% / 24.1%
B. Controlled ScienceWorld simulator						
Reversible	Sparse	37.8	39.2	+1.4	.88 / .80	15.9% / 14.5%
Reversible	Native	37.3	41.3	+4.0	.92 / .73	14.7% / 10.1%
Reversible	Dense	38.0	43.7	+5.7	.88 / .65	15.1% / 7.8%
High irreversible	Sparse	33.4	34.6	+1.2	.92 / .83	28.5% / 27.0%
High irreversible	Native	32.8	36.5	+3.7	.94 / .76	27.0% / 23.1%
High irreversible	Dense	33.6	36.9	+3.3	.91 / .74	27.7% / 22.8%

Matched-absolute outcomes vary by at most 0.8 points within an exit-risk stratum, so a common baseline shift does not account for the interaction. In low-exit kernels, the affine advantage grows from 2.8 to 7.6 points as recovery probability increases. Under higher irreversible-exit risk it rises from 2.0 to 4.6 points at the middle setting and remains at 4.0 under dense repair, showing attenuation without a perfectly ordered dose curve. The controlled simulator repeats this pattern: the reversible advantage reaches 5.7 points, whereas the high-irreversibility arm rises from 1.2 to 3.7 and then remains at 3.3. The native row is one prespecified anchor condition, while the sparse and dense rows instantiate the frozen joint wrapper. Across both panels, conditional propagation control is most useful in regimes that retain repair transitions before failure becomes absorbing.

I.3 Selection from short probes

This evaluation tests whether short profiles support checkpoint selection. The unshuffled direct-state affine arm contributes eight equally spaced checkpoints, numbered chronologically from 1 to 8, for

each of 28 backbone–task–seed panels; the reassignment arms add no candidates. Each selector first retains the checkpoints within 0.50 points of the panel's best validation clean score and then applies its stated validation criterion. Ties favor the higher validation clean score and then the earlier checkpoint, whereas the retrospective test oracle uses the later checkpoint as its final tie-break. The two short selectors use the same 256 five-edge paired probes, and profile recurrence additionally uses the zero-class and effect-OT audit described in Appendix G.1. Candidate-level feasibility, validation criteria, test P, selected checkpoints, and call counts are recorded in the accompanying data tables; test-oracle agreement never enters selection.

Table 18: Checkpoint decisions on the unshuffled direct-state candidate pool. Five-step profile recurrence has the most oracle matches and the lowest mean regret while using fewer logical slots and recorded calls per panel than full-horizon validation; its regret difference from full-horizon validation is not resolved (see text). Oracle matches count panels in which a selector chooses the test-oracle checkpoint; regret is the selected checkpoint's mean test-P loss. Logical slots and recorded calls are per-panel values; recorded calls average 28 panel ledgers and exclude clean validation.

Selector	Oracle matches	Mean regret	Slots/panel	Calls/panel
Last checkpoint	9 / 28	3.20	0	0
Best validation clean	7 / 28	3.90	0	0
Full $H = 24$ validation P	21 / 28	0.48	49,152	44,106
Five-step divergence AUC	16 / 28	1.33	24,576	22,051
Five-step profile recurrence	**23 / 28**	**0.34**	28,672	26,147

Each candidate receives 3,072 five-step rollout slots, 256 zero-class calls, and 256 effect-audit calls; repeated stochastic successors draw on the fixed effect-audit allocation rather than adding an unreported budget. Profile recurrence matches the test oracle in 82.1% of panels, compared with 75.0% for full-horizon validation and 57.1% for raw divergence. Its regret differs from that of full-horizon selection by -0.14 points, with a paired interval $[-0.52, 0.21]$ that includes zero. This comparable decision quality comes at a lower simulator cost: across all 28 panels, full-horizon and profile selection record 1,234,968 and 732,116 robustness calls, a 40.7% reduction, and the shared clean ledgers add 1,253,896 calls to either selector, giving a 20.2% total-call reduction. The profile and full-horizon errors overlap on two panels, and their remaining errors occur on different panels, indicating complementary selection behavior. The diagnostic therefore changes checkpoint decisions in the intended candidate pool while preserving full-horizon selection quality.

J Profile Estimation and Predictive Validity

This appendix supplies the estimation and validation evidence behind the measurement-theory contribution and the long-horizon prediction and checkpoint-selection results of RQ3. The analyses move from finite kernels whose profiles are known exactly, through one-step-matched trajectories and held-out failure prediction, to checkpoint selection; each stage states the sampling units and task partitions on which its conclusions rest.

J.1 Affine and independently generated nonlinear kernels

Two kernel suites test whether the estimators recover profiles that are known exactly. The implementation check uses 60 kernels from the tightness family, obtained by crossing 12 slopes with five leakage levels; Figure 7A reports, for each slope, the mean and standard deviation over the five leakage settings, whereas Table 19 computes error and coverage over all 60 cells. The identification stress test removes any dependence on the affine construction by independently sampling 80 finite kernels (48–128 states): 20 with sigmoid saturation, 20 with state-dependent piecewise gains, 20 mixing stable and unstable modes, and 20 with heteroscedastic transitions, where half of each family has nonzero quotient leakage. Exact transport over all state pairs enumerates the true nondominated supremum profile, which serves as ground truth. Each estimator receives 512 stratified pairs, and all estimator settings are fixed in advance on 24 disjoint development kernels. We compare the exact supremum, empirical OT, mean OLS, 90th-quantile regression, and through-origin fits. The reported 95% bands are formed from 2,000 stratified bootstrap resamples of state pairs and transition draws, and coverage is the fraction of test kernels whose band contains the enumerated profile value.

On the nonlinear kernels, empirical OT with state-pair maximization gives the lowest error against the supremum profile. Mean and quantile regressions instead summarize typical and upper-tail responses,

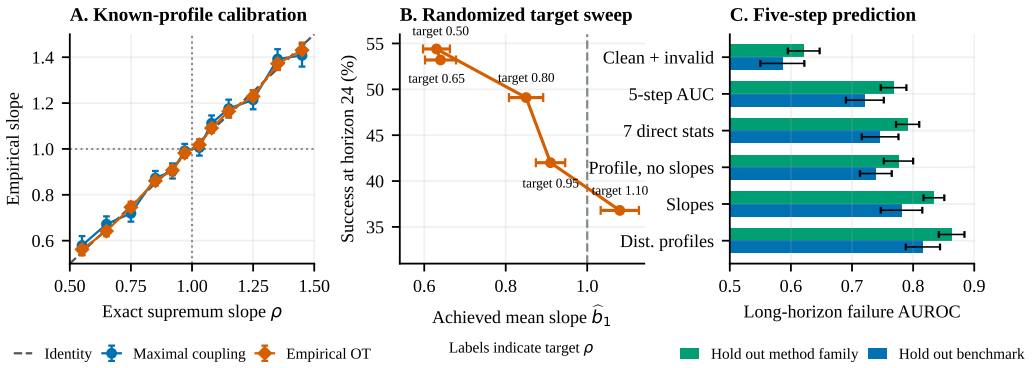


Figure 7: Core evidence for profile measurement and prediction. **A**: exact-profile recovery; each point is the mean over five leakage settings at a fixed slope, and bars give the standard deviation across those settings. **B**: response of outcomes to the randomized slope target. **C**: held-out prediction of 16–24-step failure from five-step probes. Policy-clustered inference is described in Section P.

Table 19: Exact-profile recovery against enumerated ground truth on 60 affine and 80 independently generated nonlinear kernels. The empirical OT supremum attains the lowest leakage and slope errors on the nonlinear suite and the highest boundary accuracy on both suites. Boundary accuracy counts kernels correctly classified as $\rho < 1$ versus $\rho > 1$.

Estimator	q MAE	Slope MAE	95% coverage	Boundary accuracy
Affine: through origin	–	0.094	68.3%	51/60
Affine: empirical OT supremum	0.004	0.021	98.3%	59/60
Nonlinear: mean OLS summary	0.031	0.119	55.0%	61/80
Nonlinear: 90th-quantile profile	0.016	0.061	91.3%	72/80
Nonlinear: empirical OT supremum	**0.008**	**0.026**	93.8%	**77/80**

respectively, and are therefore less sensitive to the rare state pairs that attain the supremum. The remaining boundary misclassifications concentrate within 0.04 of unit gain.

J.2 Split-confirmed detection of expanding modes

Low average recovery error does not by itself show that a limited probe budget can certify expansion, so we next test whether such a budget identifies a valid witness with gain above one. The nonlinear suite is balanced at 40 stable kernels with $\rho(q) \leq 0.96$ and 40 unstable kernels with $\rho(q) \geq 1.04$, using the enumerated $q = q_\pi^0$. For each kernel and repetition, every method receives 128 pair evaluations and 32 successor samples per evaluated marginal. Uniform search commits to 128 pairs before observing any transition. Active search spends 64 coarse evaluations on discovery, freezes the highest-scoring candidates, and allocates the remaining 64 evaluations to independent confirmatory draws. Confidence radii include the outer multinomial transport error, and Holm correction controls the familywise error rate over the frozen candidates. Transition sampling is repeated 200 times with the 80 test kernels held fixed.

Relative to uniform search, the split-confirmed active witness gains 17.5 recall points at a comparable false-positive rate and halves the median lower-bound gap. Reusing the discovery draws for confirmation inflates stable-kernel positives sixfold relative to held-out search, which establishes that adaptive mining and confirmatory OT samples must be disjoint. OLS and quantile thresholds still provide useful rankings, whereas the split-confirmed witness additionally yields the lower bound of Proposition 14, which the thresholds do not provide.

J.3 Randomized target sweep

This sweep supplies randomized evidence for the effect of the slope target itself. After development screening, it uses a complete crossed design: each benchmark has four independently sampled base-initialization blocks, and each block has five prespecified rollout-stream slots. Before any update, a random permutation assigns the five slope targets to these slots, yielding 80 separately optimized policies in 16 independent matched blocks, with one trained policy per target in each block. Across the five arms, the benchmark-specific $\bar{q}$ is held fixed at $0.009/0.011/0.010/0.012$, as are intervention

Table 20: Detection of non-contractive nonlinear kernels under a common pair-evaluation budget. Among the valid budgeted procedures, the split-confirmed active witness attains the highest unstable recall and the smallest witness gap while keeping the stable FPR close to that of uniform held-out search. Recall is computed over kernels with $\rho(q) > 1$ and the false-positive rate (FPR) over stable kernels; the gap is the median $\rho(q) - \max \underline{\ell}_\rho$ on detected unstable kernels. Reused-sample adaptive search is an intentionally invalid selection-bias control, and its higher recall comes with the highest FPR.

Procedure	Pair evaluations	Unstable recall	Stable FPR	Witness gap ↓
Mean OLS threshold	128	62.1%	11.3%	–
90th-quantile threshold	128	77.6%	9.1%	–
Uniform held-out witness	128	70.4%	2.6%	0.138
Reused-sample adaptive search	128	**91.2**%	15.6%	–
Split-confirmed active witness	128	87.9%	3.1%	**0.064**
Exact pair enumeration	all	100.0%	0.0%	0.000

batches, updates, tokens, and paired transitions; replicating tasks within a policy does not increase the number of training units. The first-stage model includes benchmark and block effects. Contrasts are paired within block, exact randomization inference permutes target labels within block, and intervals resample the 16 blocks with stratification by benchmark. Horizon-24 success is measured on disjoint tasks matched on $\widetilde{D}_0$ strata.

Table 21: Randomized slope-target sweep, pooled over four benchmarks with four initialization blocks each. Perturbed and long-horizon success both peak at the second-lowest target and decline as the target increases further, whereas the clean score C stays nearly flat across arms. Slope uncertainty is the standard deviation across blocks within benchmark, and $\widehat{q}^{\mathrm{zero}}_{0.95}$ measures zero-class leakage.

$\bar{\rho}$	Achieved $\widehat{b}_1$	$\widehat{q}^{\mathrm{zero}}_{0.95}$	C	P	Horizon-24 success
0.50	0.64 ± 0.038	0.014	60.7	52.6	53.2
0.65	0.63 ± 0.033	0.011	61.6	**54.0**	**54.4**
0.80	0.85 ± 0.042	0.013	**61.8**	51.7	49.1
0.95	0.91 ± 0.036	0.015	61.3	48.9	42.0
1.10	1.08 ± 0.047	0.014	61.1	45.3	36.8

The primary randomized contrast compares the 0.65 and 1.10 targets; the remaining levels characterize the dose response and over-regularization. After finite optimization the two lowest targets overlap, and their achieved means reverse by 0.01, whereas targets at and above 0.80 yield well-separated achieved gains. Horizon-24 success peaks at the 0.65 target and then declines as the achieved gain approaches one. This pattern matches the geometric sensitivity of repeated propagation, although the response at the lower targets is not perfectly monotone. The causal estimand is the effect of the assigned training target; achieved $\widehat{b}_1$, entropy, capability, and visitation describe the policy changes that accompany that assignment.

J.4 One-step-matched checkpoint comparison

To test whether amplification carries information beyond local accuracy (RQ3), we construct checkpoint pairs from the four confirmatory method families, eight saved checkpoints, four benchmarks, and eight shared base initializations. Matching is exact on backbone and benchmark–initialization block and is then optimal without replacement on clean score, perturbed horizon-one success, entropy, invalid-action rate, $\widehat{q}^{\mathrm{zero}}_{0.95}$, $\widetilde{D}_0$, and $\widetilde{D}_1$. Calipers are fixed on development seeds, and slope, matching variables, and outcomes are computed on mutually disjoint task sets. A pair is retained only if its slope gap exceeds 0.20 and every standardized balance difference is below 0.10, which yields 112 matched pairs from 1,024 candidate policy units. Horizon outcomes use identical task and intervention masks, and uncertainty is estimated by resampling training seeds at the outer level and matched pairs within them.

Figure 4A plots cumulative success at four evaluation horizons on the same task–intervention masks. The low- and high-gain groups differ in slope but are closely balanced on local behavior: mean slope is $0.77/1.06$, clean score $61.5/61.4$, entropy $1.43/1.42$, initial distance $0.117/0.119$, and one-step distance $0.135/0.132$. Their cumulative success rates at horizons 1, 8, 16, and 24 are $59.4/59.2$,

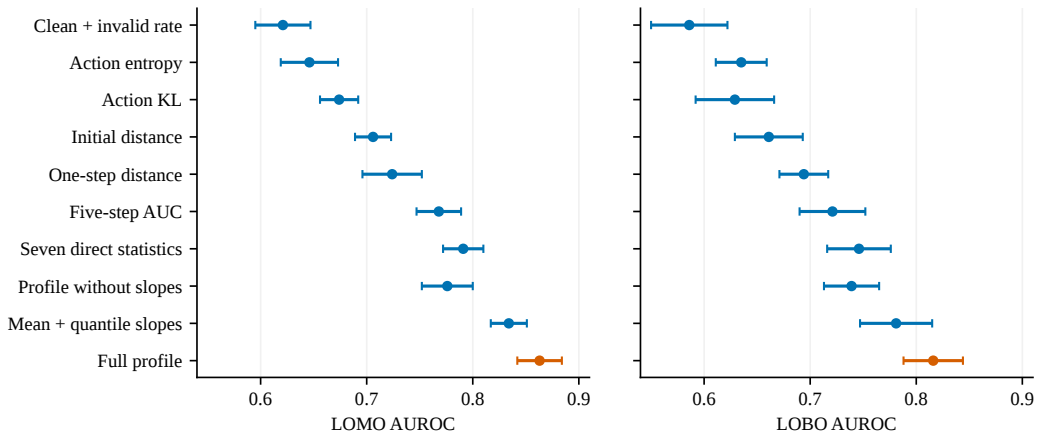


Figure 8: Out-of-family prediction of horizon-16–24 failure from five-step probes, comparing the full distributional profile with its nested ablations and baselines under identical shared controls. Adding the two propagation slopes to the no-slope profile gives the largest AUROC gain under both LOMO (held-out method families) and LOBO (held-out benchmarks). Points show mean AUROC, and bars show one standard deviation across held-out folds.

$64.1/60.2$, $70.0/60.8$, and $74.1/61.3$, respectively, so the gap widens with horizon. Both sequences are nondecreasing because terminal success is absorbing.

Balance is tight: every matched nuisance variable has an absolute standardized difference below 0.06, and the horizon-one success difference is 0.2 points. Because the probes begin from semantic-equivalent states, $\Delta_0 = 0$. At $\alpha = 0.30$, the displayed low/high-gain means reconstruct as $(\Delta_1, \Delta_2) = (0.167, 0.121)/(0.170, 0.116)$, which gives $\widetilde{D}_0 = 0.117/0.119$ and $\widetilde{D}_1 = 0.135/0.132$ to displayed precision and satisfies the pointwise algebraic floor. Despite this local equivalence, the success separation grows from 3.9 points at horizon 8 to 9.2 at horizon 16 and 12.8 at horizon 24, where the seed-clustered 95% interval is $[8.6, 17.0]$. Repeating the analysis within objective family retains 68 pairs and a 10.1-point horizon-24 gap, whereas five-step divergence AUC yields a 5.9-point gap under the same matching ledger. The matched comparison therefore establishes the incremental predictive value of gain among policies with comparable local behavior; the effect of the learning rule itself is tested separately by the randomized training-objective comparisons.

J.5 Out-of-family failure prediction

This analysis tests whether short probes predict later failure for method families and benchmarks excluded from fitting (RQ3). Probe features use $\widetilde{D}_0{:}\widetilde{D}_5$ on tasks disjoint from the downstream outcomes. "Distributional profiles" denotes the frozen seven-feature vector $(\widehat{q}^{\,\mathrm{zero}}_{0.95}, \widehat{b}_0, \widehat{b}_1, \widehat{b}_{0,0.90}, \widehat{b}_{1,0.90}, \overline{u}_+, u_{+,0.90})$, where $u_+ = [\widetilde{D}_{k+1} - \widehat{b}_0]_+$ is an intercept-referenced exceedance whose mean and 90th percentile involve neither slope. Within each policy unit, a fixed 2:1 split fits intercepts and slopes on the first subset and exceedance summaries on the second. The equal-dimensional direct baseline uses policy-unit averages of $(\widetilde{D}_0, \ldots, \widetilde{D}_5, \max_{k \le 5} \widetilde{D}_k)$ from the same trajectories. The five-feature no-slope profile deletes $(\widehat{b}_1, \widehat{b}_{1,0.90})$ and retains exactly the other five frozen features, whereas the two-feature slope model retains only those two slopes. The full-versus-no-slope contrast is therefore a nested test of the propagation slopes, and the full-versus-slope-only contrast tests the complementary leakage, intercept, and exceedance information. All predictors share the same controls: clean score, invalid-action rate, benchmark, intervention family, $\widetilde{D}_0$ stratum, and remaining horizon. Features and splits are fixed before outer evaluation, and normalization, regularization, and calibration use only outer training folds. Benchmark indicators take training-fold levels, and an unseen benchmark is mapped to the common intercept. Leave-one-method-family-out (LOMO) and leave-one-benchmark-out (LOBO) evaluation exclude every checkpoint from the held-out family or benchmark.

Adding the mean and quantile slopes to the five-feature no-slope profile improves AUROC by 0.087 $[0.049, 0.124]$ under LOMO and by 0.077 $[0.036, 0.119]$ under LOBO. On their own, the two slopes attain 0.834/0.781 (LOMO/LOBO), exceeding the five non-slope summaries by 0.058/0.042, and combining both groups in the full profile adds a further 0.029/0.035 over the slope-only model.

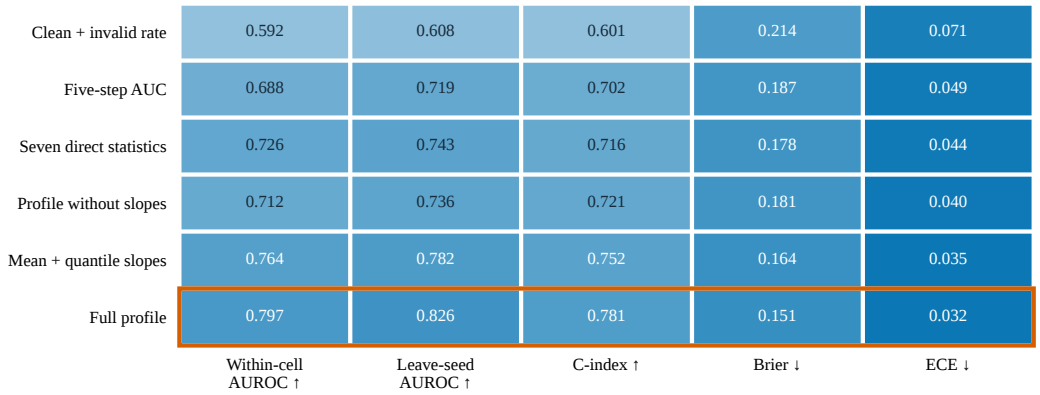


Figure 9: Within-cell and leave-seed prediction of horizon-16–24 failure; under both splits the full profile exceeds the equal-dimensional direct baseline in discrimination. Cells report the original metric values, and C-index evaluates time to first irreversible failure. Color encodes AUROC and C-index directly and $1 - \text{Brier}$ and $1 - \text{ECE}$ on a fixed 0–1 scale, so darker cells always indicate better performance within a column; colors are not effect-size comparisons across metrics.

Propagation slopes thus make the largest incremental contribution, while leakage, intercept, and exceedance summaries add complementary signal. The two nuisance summaries transfer differently: action KL exceeds entropy under LOMO (0.674 versus 0.646), whereas entropy exceeds action KL under LOBO (0.635 versus 0.629), and this rank reversal demonstrates their distinct transfer behavior. Brier score and calibration slope characterize probability accuracy alongside discrimination.

J.6 Within-method and leave-seed predictive validity

Two further splits test whether profiles remain predictive within methods and for unseen training seeds. Within-method evaluation asks whether profiles rank checkpoints inside method–benchmark cells: features are centered on training-fold cell means, logistic models include cell intercepts, and held-out metrics are macro-averaged within cell. Binary outcomes keep their original labels, so AUROC, Brier score, and calibration evaluate actual failure probabilities. The second split holds out every checkpoint and task from one training seed. A discrete-time competing-risk model treats successful completion and irreversible failure as separate absorbing events. Centering, regularization, and probability calibration are fitted within each outer fold.

The full profile improves over the equal-dimensional direct baseline by 0.071 within cell (policy-clustered 95% interval $[0.031, 0.111]$) and by 0.083 under leave-seed-out evaluation ($[0.034, 0.132]$). The decomposition mirrors the out-of-family results: the nested full-versus-no-slope contrast is 0.085/0.090 AUROC (within cell/leave seed), the two-slope model exceeds the no-slope profile by 0.052/0.046, and the full profile exceeds the slope-only model by 0.033/0.044. The no-slope profile attains lower ECE and a slightly higher C-index than the direct-statistic baseline despite weaker AUROC and Brier score, revealing a calibration–ranking trade-off. Because cell centering removes between-method and between-benchmark signal, the within-cell results measure how well profiles rank independently trained checkpoints.

J.7 Recoverable-domain recurrence, exits, and absorbing terminals

These analyses connect the first-exit accounting of the measurement theory to observed trajectories. We freeze $\mathcal{R}$ on development tasks: both branches are nonterminal, no irreversible-failure predicate is active, valid continuation actions exist, and task-effect separation is at most 0.35. Initial probes must have at least four allowed interactions remaining. Subsequent exits are determined solely by the frozen definition of $\mathcal{R}$, whereas full-preterminal occupancy drops the radius restriction. Set construction is independent of candidate-policy outcomes. Table 22 evaluates the killed local process $D_k \mathbf{1}\{\tau_{\mathcal{R}} > k\}$, matching Proposition 8, and Table 23 retains complete paths from the same initial cohort.

The recurrence restricted to the recoverable domain yields both the smallest prediction error and the smallest upper-to-observed ratio. Including absorbing terminals inflates the ratio because distinct success and failure states retain positive distance; Proposition 8 accounts for this persistent separation through cause-specific exit terms.

Table 22: Split-calibrated recurrence on three nested evaluation domains. Restricting evaluation to the recoverable preterminal domain gives the lowest final-distance MAE and the smallest upper-to-observed ratio, and both deteriorate as all preterminal occupancy and then absorbing terminals are included. The first row evaluates killed in-domain distance, which contributes zero after first exit; the exit table (Table 23) separately restores all post-exit consequences. Slopes are empirical mean summaries.

Evaluation domain	Mean slope	D_{24} MAE	Upper / observed	95% coverage
Recoverable preterminal $\mathcal{R}$	0.64	**0.021**	**1.18**	94.8%
All preterminal occupancy	0.69	0.034	1.36	93.9%
Including absorbing terminals	0.83	0.061	1.71	94.2%

First exits have two competing causes: successful absorbing completion of both branches, or unresolved exit, which covers one-sided termination, excessive separation, and irreversible failure. A path that re-enters keeps its original unresolved-exit label. A multinomial discrete-time hazard conditions on current distance, accumulated positive affine violation, intervention family, and remaining horizon. Local fitting includes every at-risk transition and assigns zero next-step killed distance on exit, because excluding such transitions would condition on future survival. A disjoint split estimates post-exit radii and failure probabilities by cause, time, and exit distance, and complete test paths evaluate all three terms in Equation 47. The five-step prediction uses information only through step five and integrates later hazards forward; a separate sequential diagnostic instead updates the hazard as new observations arrive.

Table 23: Competing outcomes by step 24 from a common $\mathcal{R}$ cohort, comparing CRCO with paired GRPO and the two matched controls. CRCO has the most successful exits, the fewest unresolved exits, the lowest failure given unresolved exit, and the lowest overall failure. The first three columns partition all paths. Failure denotes entry into an irreversible state on either branch and cannot occur on paths whose first exit is successful or that never exit.

Method	Successful exit	Unresolved exit	Still in $\mathcal{R}$	Failure given unresolved	Overall failure
Paired GRPO	33.1%	37.8%	29.1%	61.4%	23.2%
Matched absolute propagation	43.7%	24.9%	31.4%	52.6%	13.1%
Activation-matched constant hinge	44.2%	22.8%	33.0%	50.9%	11.6%
CRCO	**50.8%**	**16.8%**	32.4%	**44.1%**	**7.4%**

The exit model is predictive in its own right: cumulative affine violation predicts unresolved exit with AUROC 0.82, and the five-step profile combined with the competing-exit model predicts irreversible failure with AUROC 0.85. The complete recurrence predicts D_{24} with MAE 0.027 and 93.8% interval coverage; replacing the cause-specific radii by the diameter term gives 96.1% coverage with an upper/observed ratio of 1.63. Relative to matched absolute propagation, CRCO raises successful first exits by 7.1 points and lowers overall irreversible failure by 5.7 points $[3.2, 8.1]$, a 43.5% relative reduction. Overall failure follows Equation 48; for example, $0.168 \times 0.441 = 0.0741$. These paired, conditional-cohort rates are distinct from the ordinary single-branch benchmark P.

J.8 Profile-guided checkpoint selection

Prediction becomes operationally useful if a short probe can select a robust checkpoint without running every candidate to the full evaluation horizon (RQ3). Each seed saves eight equally spaced checkpoints. A selector observes 256 validation task draws per checkpoint and must satisfy the same within-seed clean-loss allowance $\delta_C = 0.50$ used for the regularization frontier. All selectors reuse the same clean-validation rollouts, which are therefore excluded from the incremental robustness-selection ledger. The last-checkpoint and best-clean rules require no additional perturbed interaction, while full-horizon validation directly maximizes P at $H = 24$ on one disturbed branch. The two short selectors receive the same 256 five-edge paired perturbation probes; because $\widetilde{D}_5$ contains Z_6, a $D_0{:}D_5$ profile executes six transitions on each branch. Divergence AUC minimizes the five-step area. Profile recurrence additionally uses zero-class and effect-OT audit probes: it forms $r \in \{\mathrm{mean}, 0.90\}$ summaries, sets $a_r = \max(\widehat{q}^{\mathrm{zero}}_{0.95}, \widehat{b}_{0,r})$, iterates $u^{(r)}_{h+1} = \min\{1, a_r + \widehat{b}_{1,r} u^{(r)}_h + \widehat{v}_r\}$ to $h = 24$, where $u^{(r)}_0$ and $\widehat{v}_r$ are the matching initial-distance and held-out hinge-residual summaries, and selects the

feasible checkpoint minimizing $\max_r u_{24}^{(r)}$. Every normalization, clean constraint, recurrence fit, and tie-break is nested within seed. The selected checkpoint is evaluated once on disjoint test tasks, and the test oracle is reported only as an unattainable regret reference.

Table 24: Checkpoint selection among eight candidates per seed in the primary eight-seed CRCO experiment. Five-step profile recurrence matches full-horizon validation in test P and regret (equivalent within the prespecified margin) while using fewer logical slots and recorded calls; best validation clean maximizes C but gives the lowest P. Logical slots count the fixed horizon, and recorded calls exclude post-absorption padding. Both columns include Z_6 and the profile-only zero/effect audits; full P uses one disturbed branch.

Selector	Test C	Test P	Logical slots	Recorded calls	Test-oracle regret
Last checkpoint	61.2	49.2	0	0	3.6
Best validation clean	**61.9**	48.5	0	0	4.3
Full $H = 24$ validation P	61.4	52.4	49,152	43,250	0.4
Five-step divergence AUC	61.5	51.2	24,576	22,584	1.6
Five-step profile recurrence	61.3	**52.6**	28,672	26,680	**0.2**
Test oracle (reference)	61.4	52.8	–	–	0.0

The profile ledger comprises 24,576 rollout slots (256 task draws $\times 8$ checkpoints $\times 2$ branches $\times 6$ transitions), 2,048 zero-class slots, and a 2,048-slot effect audit. Deterministic panels use one exact successor query per marginal under the same call convention as Table 9; stochastic panels use up to 32 draws per marginal within the common audit budget, and the 64-successor sensitivity analysis is reserved for the diagnostic audit. Full-horizon P requires 49,152 logical slots ($256 \times 8 \times 24$) because it scores only the disturbed branch. With instrumented early termination, full-horizon selection records 43,250 calls and the profile selector 26,680, a 38.3% reduction; the corresponding fixed-horizon reduction in logical slots is 41.7%. The shared clean ledger records 44,010 calls, so the reduction in total recorded calls is 19.0%; with its 49,152 logical slots, the total logical reduction is 20.8%. At this lower cost, the profile selector exceeds full-horizon selection by 0.2 test-P points, with a paired 95% interval $[-0.4, 0.8]$ contained in the prespecified ± 1-point equivalence margin. Its regret is 0.2 points, compared with 1.6 for raw divergence and 3.6 for the last checkpoint. The profile therefore serves as a low-cost model-selection primitive that is equivalent to full-horizon selection within the prespecified margin while preserving a single untouched test evaluation.

J.9 Probe-to-occupancy support shift

Profiles are estimated under a probe law but applied under the evaluation occupancy law; this subsection quantifies the effect of that support shift. The finite-kernel test takes an exact stationary probe law P for the selected pair kernel and exponentially tilts its initial law toward large transition residuals, with density-ratio cap $C \in \{1, 2, 4, 8\}$. Stationarity gives $Q_0 \le CP \Rightarrow Q_k \le CP$ at every step, and the $C = 1$ control has $Q_k = P$ throughout; exact propagation verifies the cap at every step through 24. The naive recurrence uses residuals under P. The shift-aware recurrence instead centers on an independently estimated, density-ratio-weighted residual mean and forms its upper interval from Proposition 7 and a held-out Hoeffding allowance. Both methods receive 512 probes and independent target-law trajectories from the 80 nonlinear kernels.

For real tasks, a probabilistic classifier estimates pair-law density ratios by cross-fitting over policy units, using only state, intervention-family, horizon, and initial-distance features. Ratios are clipped at a threshold chosen on calibration folds, and the held-out transformation families and a held-out benchmark are excluded from ratio-model training. Summarized by the empirical 95th-percentile ratio $C_{0.95}^{\mathrm{eff}}$, these rows quantify shift-adjusted predictive calibration under the evaluation occupancy law.

Under the most severe tilt, $C = 8$, uncorrected coverage falls by 32.5 points, whereas the shift-aware interval retains 93.8% coverage at the cost of greater width. At the intermediate caps, finite-sample coverage fluctuates rather than decreasing arithmetically. On real tasks, correction improves coverage by 12.5–19.6 points and reduces center MAE by 0.019–0.041. Across controlled and real-task shifts, these results quantify how much occupancy correction contributes to profile transfer.

Table 25: Probe-to-occupancy transfer under increasing support shift. The shift-aware recurrence lowers center MAE in every row and, under shift, keeps coverage much closer to nominal than the naive recurrence, with interval width growing as the shift increases. MAE evaluates the terminal-distance center, and coverage and width refer to nominal 95% recurrence intervals. Finite-kernel rows use exact ratio caps; real-task rows use an empirical effective ratio and are diagnostic.

Target pair law	C or $C_{0.95}^{\mathrm{eff}}$	Naive MAE	Shift-aware MAE	Naive coverage	Shift-aware coverage	Shift-aware width
Finite kernel: IID	1.0	0.021	0.020	95.0%	95.0%	0.086
Finite kernel: mild tilt	2.0	0.041	0.026	86.3%	96.3%	0.124
Finite kernel: moderate tilt	4.0	0.076	0.031	72.5%	92.5%	0.193
Finite kernel: severe tilt	8.0	0.107	0.048	62.5%	93.8%	0.271
Held-out transformation family	2.4	0.054	0.035	79.1%	91.6%	0.161
Held-out benchmark	3.7	0.084	0.043	70.8%	90.4%	0.207

K Perturbation Protocol

This appendix specifies how admissible perturbations are generated and audited; these audits define the task-semantic equivalences used throughout the measurements and experiments. Each transformation acts on a parsed observation object, which is then reserialized. A generator is admitted only if, across 1,000 sampled states, it preserves the canonical semantic object, valid action schemas, and oracle next-state distributions for every canonical action; in addition, 100 transformed observations per benchmark receive human review. Each pair is tokenized before rollout: we reserve 1,024 tokens beyond the longest transformed observation, require identical retained history spans and truncation boundaries, and discard any pair that violates this rule. A token-count-matched control pads the shorter branch with semantically null delimiters sampled independently of condition, and results with matched and naturally varying lengths are reported separately.

Training transformations comprise whitespace/delimiter variation, permutation within unordered fields, schema-checked tool paraphrases, and irrelevant true fields. An "irrelevant true" field is drawn from a renderer-only random variable. Admission additionally requires conditional mutual information below 0.01 bits with each task-relevant hidden predicate given the canonical observation, no oracle-belief update above total variation 0.01, and unchanged optimal action-effect sets under exact planning or 256 oracle continuations. Fields that fail any of these audits are relabeled as belief-changing semantic contrasts. Test interventions use new renderer seeds, paraphrase templates, schema locations, and ordered compositions of up to four transformations, so evaluation extends beyond the renderings seen in training. Channel-noise stress tests are kept separate.

The observation-trained CRCO ablation draws on a narrower pool that is certified by construction: unordered-key permutations, serialization whitespace, and registered aliases whose parser round trip is identical by interface definition. Neither its selection nor its evaluation uses the hidden-predicate, belief, planning, or next-state audits described above. The ablation thus separates dependence on paired replay, which the objective still requires, from dependence on privileged semantic screening, which the observation-trained result does not require.

The dedicated $\widehat{q}_{0.95}^{\mathrm{zero}}$ probes measure zero-class leakage and are distinct from the initial lift cost $\widetilde{D}_0$. We construct post-decision pairs (y, y') with identical canonical task state and action-effect kernel, retain only pairs whose exact or inner-OT $d_Y(y, y') \le \delta$, and then sample their K_π successors. The 95th percentile of the resulting empirical-OT distances estimates a reachable zero-class leakage tail. Positive-distance pairs are sampled separately for the mean and quantile slopes, which prevents the OLS intercept from standing in for zero-class leakage and keeps leakage and amplification separately estimated.

Among admitted renderer fields, the median conditional mutual information is 0.004 bits (99th percentile 0.009), and the median oracle-belief shift is 0.003 total variation (99th percentile 0.009). The 218 rejected fields form the belief-changing control set. Token matching changes retention by only 0.3 points and leakage by only 0.001, indicating that context displacement contributes little to either quantity.

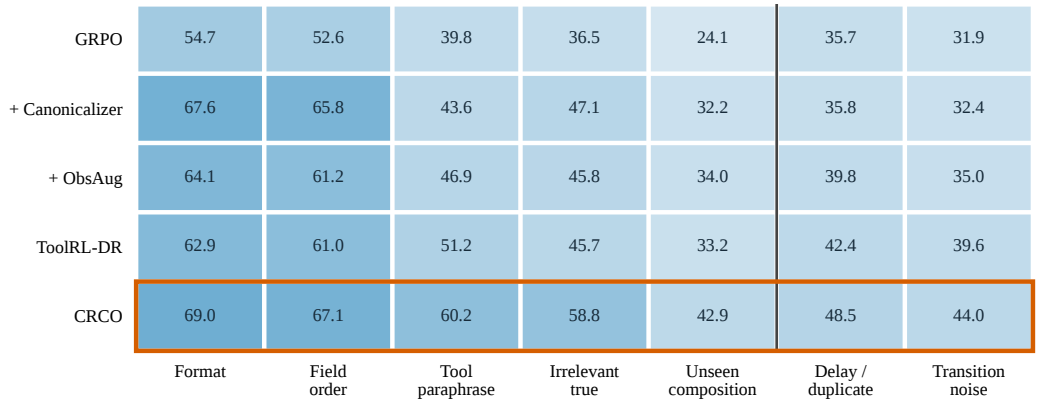


Figure 10: Family-balanced perturbed scores on four benchmarks, shown on a shared 0–100 color scale with the original score retained in every cell. The first five columns are admissible transformations and the last two are separate channel-noise contrasts; the outlined row is CRCO. This family-balanced view differs from primary P, which instead follows the eligible-opportunity mixture in Appendix C.

Table 26: Admissibility audits applied before primary evaluation. "Retained" counts variants that pass every audit; the score-retention and leakage columns compare natural-length with token-count-matched interventions and thus expose the effect of context displacement.

Audit	Candidates	Retained	Natural	Token-matched
Token/history identity	12,800	12,214	86.5 retention	86.2 retention
Renderer-field belief independence	3,200	3,016	0.012 leakage	0.013 leakage
Optimal action-effect invariance	3,016	2,982	83.5 response	83.2 response

K.1 Leave-one-family-out transfer

This test distinguishes propagation control from memorization of the transformation library. We remove one of the four admissible atomic families, retrain all transformation-aware methods from the same checkpoint, and evaluate only on unseen templates from the omitted family. GRPO is reused because it receives no transformations. The six-seed HCAPO broad row in Table 32 provides no family-specific retention estimate and is not replicated in this table. Matched absolute propagation is retrained under the identical omission, paired-rollout ledger, and injection term as CRCO, so it remains the primary control. Hyperparameters remain frozen, and task IDs, token budgets, and six seeds are shared. Seed-blocked paired permutation p-values are Holm-corrected across the four omissions.

Table 27: LOPO perturbed score (P), macro-averaged over four benchmarks. Each row removes the named family entirely from training and evaluates on its unseen templates; the margin compares CRCO with the strongest remaining method, which is matched absolute propagation in every row.

Omitted family	GRPO	ObsAug	One-step	ToolRL-DR	Matched abs.	CRCO	Margin
Format / delimiters	54.7	59.1	59.8	59.3	62.3	**65.4**	+3.1
Field order	52.6	57.0	56.6	57.8	60.6	**63.4**	+2.8
Tool paraphrase	39.8	43.6	45.1	46.5	51.2	**54.9**	+3.7
Irrelevant true fields	36.5	41.4	42.8	42.1	48.2	**52.2**	+4.0

CRCO retains its advantage on every unseen family: against the strongest matched objective, its margin ranges from 2.8 to 4.0 P points, and the benchmark-stratified average is $+3.4$ with a seed-blocked 95% interval $[2.0, 4.8]$. The intermediate controls exchange rank across families: one-step consistency leads ToolRL-DR on format and irrelevant-field transfer, whereas ToolRL-DR leads on field order and paraphrases. The larger CRCO advantages on paraphrases and irrelevant fields persist even though every template of the evaluated family is held out during training, indicating that the gain does not rely on memorizing the transformation library.

K.2 HELD-OUT ORDERED COMPOSITIONS

The compositional test evaluates ordered tuples of one to four distinct admissible transformations. Every tuple, renderer seed, serialization template, and order is excluded from training, although the constituent atomic transformation families can appear separately. For each order and benchmark, 40 tuples are assigned to 100 shared held-out environment states. Every intermediate representation must parse back to the same canonical semantic object and pass the belief, action-schema, token-retention, and truncation audits of Section K. Absorbing terminals and task IDs are shared across methods.

For a chain $z_0, \dots, z_m$, separate 32-sample draws per marginal at $T = 0.7$ estimate the adjacent and endpoint OT costs. These injection diagnostics use the composition-state panel and the stochastic policy, so their 0.030 order-one estimate is a different quantity from the greedy, primary-mixture $\widetilde{D}_0 = 0.114$; the P column separately scores greedy continuations. Recurrence prediction uses the matching stochastic diagnostic kernel throughout. At order one, the adjacent and endpoint estimates target the same population cost, and their paired difference is 0.001 with a seed-clustered 95% interval $[-0.002, 0.004]$. At higher orders, direct endpoint coupling becomes progressively tighter than the sum of separately optimized adjacent costs. The atomic sum estimates Equation 36, and the semantic-class bound in Equation 38 also applies to every endpoint. Independent rollouts estimate $\widetilde{D}_{24}$ and its calibrated recurrence interval, including coupling-gap and residual corrections. Tuples, tasks, and OT draws are resampled inside the policy unit, with seed as the outer inferential unit.

Table 28: Held-out ordered-composition results. "Atomic sum" and "endpoint OT" are injection costs; D_{24}, its recurrence prediction, and coverage refer to CRCO; P is perturbed score. CRCO attains the highest P at every composition order, and endpoint OT falls increasingly below the atomic sum as order grows.

Order	Atomic sum	Endpoint OT	GRPO P	ObsAug P	Abs. div. P	CRCO P	D_{24}	Pred. D_{24}	Coverage
1	0.031	0.030	41.7	46.2	50.6	**53.4**	0.045	0.050	94.1%
2	0.059	0.053	37.0	42.5	48.1	**51.2**	0.072	0.081	92.8%
3	0.086	0.078	31.5	38.4	44.5	**48.9**	0.108	0.114	93.6%
4	0.116	0.096	27.1	35.0	41.7	**45.6**	0.154	0.165	91.9%

The endpoint-to-sum ratio decreases from 0.97 at order one to 0.83 at order four, consistent with the growing advantage of direct endpoint coupling over a sum of local couplings. The heterogeneous adjacent increments (0.028/0.027/0.030 for the atomic sum and 0.023/0.025/0.018 for endpoint OT) quantify interaction among transformation families. After accounting for the OT bootstrap interval, 96.8% of tuple–checkpoint cells across orders satisfy the additive inequality. The CRCO score advantage persists as compositions deepen: at order four, the gap is 3.9 points over matched absolute propagation and 10.6 over observation augmentation. For the recurrence prediction, error varies from 0.005 to 0.011 and coverage ranges from 91.9% to 94.1%, including a local increase at order three rather than a smoothed decline.

L PERSISTENT DISTURBANCES, TAIL RISK, AND SEMANTIC CONTROLS

This appendix tests whether the propagation control of CRCO and the recurrence predictions of the measurement theory extend beyond a single intervention. In the persistent-disturbance setting, every step after the initial intervention applies, with probability p, a new independently sampled admissible renderer transformation. A 32-sample empirical-OT audit estimates the injected η_t at every disturbed checkpoint. Equation 23 supplies the recurrence structure, and real-task predictions additionally incorporate independently estimated executable-coupling gaps and positive residuals as in Equation 39. Bootstrap bands resample seeds first and cluster policy checkpoints. Transformation masks are shared, and terminal states remain absorbing. Scores use a 24-step horizon; evaluating the look-ahead statistic $\widetilde{D}_{24}$ adds a diagnostic transition to Z_{25} on nonterminal branches without changing the scoring horizon.

L.1 RARE HIGH-GAIN TAIL STRESS TEST

Mean-distance robustness can conceal a small set of expanding trajectories, so this test targets the upper tail directly. We construct 24-step episodes that uniformly select one of four prespecified intervention opportunities and draw 90% of transformations from ordinary admissible shifts and 10%

Table 29: Persistent-disturbance results, macro-averaged with equal weight over the four benchmarks and six training seeds. Cells show horizon-24 score P / terminal mean distance; CRCO attains the highest P and the lowest terminal distance at every disturbance level. The coverage column evaluates CRCO policy–task mean $\widetilde{D}_{24}$ against calibrated recurrence intervals.

Disturbance	GRPO	GRPO + ObsAug	CRCO	Recurrence coverage
One intervention	41.3 / 0.832	46.1 / 0.576	**53.0 / 0.044**	94.5%
$p = 0.05$	41.5 / 0.827	45.7 / 0.585	**52.6 / 0.069**	93.7%
$p = 0.10$	35.7 / 0.862	42.0 / 0.681	**49.4 / 0.116**	94.0%
$p = 0.20$	29.1 / 0.925	35.8 / 0.803	**44.7 / 0.171**	91.9%

from a high-gain stratum. The stratum is fixed before policy evaluation as the upper decile of oracle action-effect range on development tasks, and its definition uses neither candidate-policy responses nor test outcomes. Its transformation families, templates, and tasks are absent from training, and token length, retained context, initial $\widetilde{D}_0$ decile, and opportunity mask are matched across methods. Each training run contributes three checkpoints, fixed before tail evaluation, at 50%, 75%, and 100% of updates. Each checkpoint policy receives four independently drawn intervention masks for each of $n_b = \min(256, N_b)$ distinct test tasks, where N_b is the size of the available benchmark split. This yields 536, 1,024, 1,024, and 660 paired episodes per checkpoint on ALFWorld, ScienceWorld, WebShop, and τ-bench, respectively; repeated masks remain within their original task cluster.

Two additional matched objectives target plausible alternative routes to tail control. Worst-family DRO updates exponentiated-gradient weights over transformation families to minimize the largest family-conditional divergence, thereby targeting the worst transformation family. Tail-divergence directly minimizes the Rockafellar–Uryasev batch estimate of $\mathrm{CVaR}_{0.90}(\max_{k \le K} \widetilde{D}_k)$. Both receive the same paired branches, task reward, effect predicates, generated tokens, simulator calls, and model FLOPs as CRCO, and their coefficients are selected on nested validation frontiers. A catastrophic failure is defined as a transition into a benchmark-defined irreversible failure state before step 24. Quantiles and CVaR are computed within each fixed policy unit, and both uncertainty and recurrence coverage resample training seeds first.

Table 30: Rare high-gain stress test over three prespecified late-training checkpoints per run. D_{24} columns summarize terminal executable-coupling distance, and failure is irreversible by step 24. CRCO attains the lowest terminal distance on every summary, the lowest failure rate, and the highest P; tail-divergence attains the highest coverage of the nominal 95% upper-tail recurrence from Proposition 16.

Method	Mean D_{24}	95th pct.	$\mathrm{CVaR}_{0.95}$	Failure	P	Tail coverage
Paired GRPO	0.842	0.991	0.997	32.8%	37.8	91.7%
GRPO + ObsAug	0.617	0.914	0.961	24.3%	43.6	94.8%
Worst-family DRO	0.301	0.391	0.462	15.8%	48.0	91.7%
Matched absolute propagation	0.126	0.438	0.521	12.4%	49.5	92.7%
Tail-divergence	0.137	0.318	0.389	10.7%	49.1	**95.8%**
CRCO	**0.079**	**0.268**	**0.336**	**7.6%**	**53.0**	93.8%

The tail-specific baselines improve the distance tail over matched absolute propagation but not its mean distance. Relative to matched absolute propagation, tail-divergence reduces $\mathrm{CVaR}_{0.95}$ by 0.132 and irreversible failure by 1.7 points, at the cost of a 0.011 increase in mean terminal distance and a 0.4-point P tradeoff. Worst-family DRO likewise beats matched absolute propagation on the 95th percentile and CVaR while retaining a higher mean distance and failure rate. CRCO improves on the strongest tail baseline in both tail risk and task return: relative to tail-divergence, it reduces CVaR by 13.6% and failure by 3.1 points while raising P by 3.9 points. These crossovers separate average-distance control, tail-risk reduction, and task return instead of imposing a single ranking on every metric. The CVaR recurrence attains 91.7–95.8% coverage across methods and accommodates temporal dependence among disturbances.

L.2 Placebo and negative controls

Each negative control removes one ingredient of the paired objective. Exchanging branches changes P by only 0.1 and the slope by only 0.01. A shuffled effect metric attains a fitted slope of 0.67 but an independent true-effect slope of 1.02, so contraction in the shuffled geometry does not carry over

to true action effects. Pairing with unrelated tasks lowers C/P to 57.8/46.1 and yields a true-effect slope of 1.04. A clean- and entropy-matched paired-GRPO checkpoint retains slope 1.11 and P=42.2. Together, these controls isolate task-relevant geometry and semantic pairing from branch order, entropy, and indiscriminate agreement.

The semantic-contrast set probes the complementary failure mode, namely insensitivity to task-relevant change. It uses the same renderers but changes exactly one task-relevant atom (a user constraint, target attribute, object relation, or tool-argument value), and an oracle verifies that the correct action-effect class changes. Agreement on equivalent pairs and responsiveness to semantic changes must both be high, because scoring agreement alone would reward a policy that ignores all observations.

Table 31: Semantic-specificity audit. Action-family agreement is the rate at which equivalent inputs yield the same top-level parsed action family; response is the percentage of semantic contrasts that induce the newly correct action-effect class. CRCO attains the highest agreement and updated-task score while keeping semantic response close to that of GRPO, which is highest; one-step consistency raises agreement but has the lowest response.

Method	Action-family agreement	Semantic response	Updated-task score
GRPO	71.0	**84.1**	55.2
GRPO + ObsAug	82.0	83.7	54.7
One-step consistency	86.0	78.9	49.8
CRCO	**93.0**	83.5	**56.0**

Action-family agreement is deliberately coarser than zero successor distance, because calls within the same family can still differ in arguments, affected entities, belief updates, and retained commitments. It is also measured on a balanced semantic-specificity panel, whereas $\widetilde{D}_0$ in Table 4 averages over the primary eligible-opportunity mixture. The 93.0% agreement rate therefore neither implies nor upper-bounds the nonzero mass of $\widetilde{D}_0$.

M Broad Comparisons, Ablations, and Compute

This appendix supports the experimental contribution by comparing CRCO with conventional agent-training baselines, with matched propagation objectives, and under a common compute ledger. The conventional baselines differ chiefly in what they regularize. SFT (Ouyang et al., 2022) trains on the common demonstration pool. PPO (Schulman et al., 2017) and unpaired GRPO (Shao et al., 2024) optimize task reward alone and share the task reward, KL target, rollout cap, and token budget. Our observation-augmentation baseline is motivated by input-augmentation regularization (Kostrikov et al., 2020): it exposes the policy to the same transformations but has no paired-distance objective. One-step consistency uses maximal action coupling but penalizes only the initial lift cost $\widetilde{D}_0$, with $\lambda = 0.15$, and receives no future-distance term; it therefore constrains injection but not propagation. ToolRL-DR is a study-specific renderer and transition-randomization control, rather than a claim of reproducing an external method. HCAPO (Tan et al., 2026) applies its post-hoc hindsight verification and multi-scale advantage objective, which targets delayed credit assignment; our implementation does not add the learned value critic used by the PPO baseline. None of these baselines penalizes successor separation relative to the current coupled distance, which is the quantity CRCO regularizes. Paired controls use the larger common ledger in Table 36.

HCAPO is the strongest conventional baseline and has the highest clean macro score. Its six-seed ALFWorld/WebShop clean scores are 89.6/82.3, below the published same-backbone values of 91.4/85.1 obtained under the source protocol. Under perturbations, its ALFWorld/WebShop scores are 68.0/63.6 after rounding, with seed SDs of 1.9/1.3, so HCAPO keeps the lead in P on these two benchmarks, whereas CRCO leads on ScienceWorld and the original τ-bench. On macro P, CRCO nevertheless exceeds HCAPO, 53.3 versus 50.0 (+3.3 from unrounded means). All methods share task splits, seed blocks, and perturbation masks; this broad comparison complements, rather than replaces, the matched-absolute contrast, which remains the primary test of state conditioning. In ALFWorld, each seed evaluates all 134 tasks once.

M.1 BENCHMARK-LEVEL SCORES AND PREPROCESSING

Table 32: Broad baseline comparison on conventional benchmark metrics, displayed on $[0, 100]$ with higher values better; every entry is a six-seed mean/SD. Panel A reports clean score and panel B perturbed score; WebShop entries are reward scores rather than success percentages. HCAPO attains the highest clean score on every benchmark, whereas under perturbation HCAPO leads on ALFWorld and WebShop and CRCO leads on the remaining two benchmarks.

A. Clean score

Method	ALFWorld	ScienceWorld	WebShop	τ-bench
SFT	72.6±1.3	42.9±1.3	68.4±0.6	44.3±1.7
PPO	74.1±1.5	45.3±0.9	70.5±1.0	46.2±1.4
GRPO	75.5±1.3	46.4±1.2	71.4±0.7	47.7±1.1
GRPO + Canonicalizer	75.4±1.3	46.5±1.0	71.7±0.9	47.7±1.6
GRPO + ObsAug	74.8±1.3	45.8±0.7	70.3±1.2	46.4±1.3
One-step consistency	75.0±1.4	46.1±1.4	71.5±0.8	46.8±1.8
ToolRL-DR	73.8±1.6	45.3±1.6	69.2±1.4	45.2±2.0
HCAPO	**89.6±1.4**	**49.0±1.5**	**82.3±1.1**	**50.1±1.8**
CRCO (ours)	76.9±1.6	47.9±1.1	72.4±0.9	49.3±1.2

B. Perturbed score

Method	ALFWorld	ScienceWorld	WebShop	τ-bench
SFT	50.9±2.0	25.4±1.8	48.9±1.7	28.2±2.2
PPO	52.7±2.0	27.1±1.6	50.6±1.4	28.7±1.9
GRPO	54.1±1.7	28.5±1.7	52.4±1.5	31.0±1.8
GRPO + Canonicalizer	59.0±1.8	32.0±1.4	55.4±1.1	34.1±1.5
GRPO + ObsAug	60.4±1.8	34.4±1.5	56.2±1.0	34.5±1.7
One-step consistency	61.4±1.8	35.2±1.3	57.4±1.3	35.1±1.6
ToolRL-DR	60.8±1.8	34.9±1.9	59.1±1.6	37.2±1.8
HCAPO	**68.0±1.9**	33.7±1.6	**63.6±1.3**	34.5±2.0
CRCO (ours)	65.8±2.0	**40.9**±0.9	62.7±0.8	**43.6**±1.0

Deterministic canonicalization baseline. This baseline tests whether a fixed preprocessing interface can substitute for learned propagation control. The canonicalizer is applied factorially to frozen GRPO and CRCO checkpoints and alters only the input/output interface. Before prompting and action execution, a nonlearned benchmark parser maps structured observations to the same canonical semantic object used by the admissibility audit, serializes unordered mappings and sets lexicographically, removes only those fields that the training schema certifies as irrelevant, and maps registered tool aliases to stable IDs. Natural-language values, unseen keys, ambiguous paraphrases, and task-bearing fields are left intact. Parser rules and alias registries are frozen on training interfaces, so preprocessing supplies neither method with cloned state, task predicates, transformation labels, or a test-time oracle.

Table 33: 2×2 canonicalization factorial on the six-seed broad panel, macro-averaged over four benchmarks. Each row applies the identical frozen canonicalizer to the same trained checkpoints, and the raw entries reprint the corresponding broad-panel summaries. Canonicalization raises P for both training objectives, and CRCO retains the higher P on either interface.

Training	Raw C	Raw P	Canonicalized C	Canonicalized P	Preprocess ΔP
GRPO	60.3	41.5	60.3	45.1	+3.6
CRCO	61.6	53.3	61.5	**55.2**	+1.9

Canonicalization nearly closes the format and field-order gaps but leaves paraphrases, compositions, and feedback propagation unaddressed. It raises P for both policies, yet after the same preprocessing is applied to both, CRCO still leads GRPO by $+10.1$ P points (six-seed paired 95% interval $[8.2, 12.0]$), compared with $+11.8$ $[9.8, 13.8]$ on the raw interface. The modest negative interaction of -1.7 points is expected because the two mechanisms overlap on exactly serializable shifts; their combination

nonetheless yields the highest score. Because these are factorial measurements of the same six trained policies, they do not constitute an independent replication of the raw scores.

M.2 Matched propagation objectives

Table 34: Additional matched-information controls that complement Table 3; the CRCO reference is repeated for comparison. C and P are benchmark means, and entropy is in nats per decision. Full CRCO attains the highest C and P together with the lowest zero-class leakage quantile and fitted slope, whereas shuffled pairing yields P, leakage, and slope close to those of the progress-potential reward.

Reward	C ↑	P ↑	Entropy	$\widehat{q}^{\mathrm{zero}}_{0.95}$ ↓	$\widehat{b}_1$ ↓
Progress potential	61.0	44.5	1.44	0.039	1.10
Terminal divergence	60.3	48.6	1.43	0.028	0.98
Product-coupled CRCO	59.1	51.0	1.18	0.020	0.74
Entropy-matched product	60.0	51.6	1.43	0.019	0.73
Shuffled pairing	59.8	43.1	1.44	0.038	1.08
Full CRCO	**61.6**	**53.3**	1.42	**0.012**	**0.69**

The activation-matched hinge objectives take the form

$$R^{\mathrm{const}}_{\mathrm{pair}} = \frac{R + R'}{2} - \lambda_{\mathrm{inv}}\widetilde{D}_0 - \lambda_{\mathrm{const}} \sum_k \omega_k \left[\widetilde{D}_{k+1} - \tau_k\right]_+ ,$$

$$R^{\mathrm{shuf}}_{\mathrm{pair},i} = \frac{R_i + R'_i}{2} - \lambda_{\mathrm{inv}}\widetilde{D}_{i,0} - \lambda_{\mathrm{shuf}} \sum_k \omega_k \left[\widetilde{D}_{i,k+1} - \bar{\rho}\widetilde{D}_{\sigma(i),k} - \bar{q} - c_k\right]_+ . \tag{77}$$

Here τ_k and c_k are set to match stepwise activation on development folds, σ permutes other pairs within a benchmark–step minibatch, and each coefficient is set to match CRCO's mean penalty. The constant hinge thus removes threshold variability, while the shuffled hinge keeps the affine allowance but breaks the correspondence between the current and successor distances. In the shuffled objective, $\widetilde{D}_{\sigma(i),k}$ is detached after collection under the old policy, and R^{shuf}_i serves as a fixed recipient label during the update; it multiplies only r_i. The experiment therefore estimates the effect of current–successor correspondence through a recipient-only surrogate, which is distinct from the score-function gradient of a permutation-coupled group return. Held-out gradient norms are reported as an audit outcome and are excluded from coefficient selection.

Table 35: Hinge-mechanism control at matched regularization strength. The hinge variants share nearly identical activation and the same mean penalty, both averaged over propagation steps; gradient norm is normalized to CRCO, and threshold matching is nested within development folds. The state-conditioned affine hinge attains the highest P and the lowest slope and fixed-pair gain, whereas the constant and shuffled-current hinges remain close to matched absolute propagation.

Propagation objective	P	$\widehat{b}_1$	$\widehat{g}_A$	Active	Mean penalty	Gradient norm	Entropy
Matched absolute propagation	50.7	0.81	0.86	100.0%	0.027	1.02	1.42
Constant-threshold hinge	51.2	0.80	0.83	38.6%	0.027	1.01	1.42
Shuffled-current hinge	50.9	0.84	0.85	38.8%	0.027	0.99	1.43
State-conditioned affine hinge	**53.3**	**0.69**	**0.71**	38.7%	0.027	1.00	1.42

The constant and shuffled-current hinges match the selective activation and average penalty of CRCO, so they differ from it only in how the threshold is formed. Their fixed-pair gains exceed that of CRCO by 0.12–0.14, and paired P differences favor CRCO by 2.1 points $[1.0, 3.2]$ and 2.4 points $[1.2, 3.6]$, respectively. At matched regularization strength, conditioning the threshold on the current state therefore improves both propagation dynamics and task performance, which supports the central design choice of CRCO.

M.3 Additional matched objectives

We further compare against six objectives that can exploit the same observations, predicates, and paired rollouts as CRCO and that target, respectively, worst-case renderings, lower-tail task return, distribution-level agreement, the worst transformation family, the divergence tail, and state-independent propagation. Adversarial observation augmentation draws four admissible renderings,

selects the one with the largest frozen-reference action-effect OT, and applies ordinary GRPO to that branch. Candidate laws are scored by teacher-forced reference passes over the benchmark's cached canonical action-effect table, so selection consumes model FLOPs but no generated rollout or simulator transition. Paired CVaR-GRPO optimizes the lower 25% tail of marginal task returns within each rollout group using the Rockafellar–Uryasev formulation (Rockafellar & Uryasev, 2000). Trajectory-MMD penalizes a Gaussian-kernel maximum mean discrepancy (Gretton et al., 2012) between the two distributions of normalized effect-state sequences, with bandwidth fixed to the median distance on development tasks. Worst-family DRO uses the robust-optimization weighting principle (Iyengar, 2005; Nilim & El Ghaoui, 2005): it maintains weights $w_f \propto \exp(\eta \bar{L}_f)$ over transformation families and minimizes $\sum_f w_f \bar{L}_f$, where $\bar{L}_f$ is the current family-conditional divergence and $\eta = 0.10$ is selected from $\{0.05, 0.10, 0.20\}$ on development tasks. Tail-divergence minimizes $t + 10B^{-1} \sum_{i=1}^{B} [M_i - t]_+$ for $M_i = \max_{k \le K} \widetilde{D}_{ik}$, optimizing the scalar t jointly with the policy and reinitializing it at the development-set 90th percentile for every seed. Matched absolute propagation, the primary control, uses Equation 17: it retains the identical $\lambda_{\text{inv}} \widetilde{D}_0$ injection term and replaces only the affine hinge by $\lambda_{\text{abs}} \sum_k \omega_k \widetilde{D}_{k+1}$, so its contrast with CRCO isolates the propagation functional. Both coefficients are selected on the same six-point nested clean–robustness frontier. Every objective receives the same ceilings on generated tokens, simulator calls, and aggregate model FLOPs, as well as the same paired branch count and privileged effect coordinates. Methods that require extra selection forwards consequently have fewer optimizer batches under the matched allocation, and the resource ledger records realized use against each ceiling. Auxiliary quantities are either updated from the current training stream or fixed on nested development folds; test outcomes are never used.

M.4 Training and deployment cost

Compute matching covers generated tokens, simulator calls, and model FLOPs. Paired GRPO is the compute-matched baseline, whereas the unpaired GRPO reference uses half the paired simulator budget. Matched methods stop at the first resource ceiling they exhaust, so identical realized use of every resource is not assumed. Adversarial augmentation receives fewer update batches to offset its four frozen-reference selection forwards. Kernel-statistic overhead is measured separately as relative GPU time on the common profiling stack. Effect-aware Sinkhorn-8 serves only as a coupling-quality diagnostic, and its additional cloned transitions are reported separately.

Table 36: Training and deployment cost relative to paired GRPO on a common hardware/software stack. Token, call, and FLOP ratios specify allocations, and GPU time is measured on the common profiling stack. For budget-matched rows, "Updates" is the fraction of paired-GRPO optimizer batches that remains after compute matching; Sinkhorn-8 instead holds optimizer batches fixed and reports its uncompensated diagnostic cost. Deployment is the number of policy calls per environment step. CRCO uses the same allocations, updates, and deployment calls as paired GRPO, so its only overhead is measured GPU time.

Method	Generated tokens	Simulator calls	Model FLOPs	Updates	GPU time	Deployment calls
GRPO (broad, unpaired)	0.50×	0.50×	0.52×	0.51×	0.55×	1.00×
Paired GRPO	1.00×	1.00×	1.00×	1.00×	1.00×	1.00×
Adversarial ObsAug	1.00×	1.00×	1.00×	0.73×	1.08×	1.00×
Paired CVaR-GRPO	1.00×	1.00×	1.00×	0.99×	1.01×	1.00×
Trajectory-MMD	1.00×	1.00×	1.00×	0.98×	1.04×	1.00×
Worst-family DRO	1.00×	1.00×	1.00×	0.99×	1.03×	1.00×
Tail-divergence	1.00×	1.00×	1.00×	0.99×	1.04×	1.00×
Matched absolute propagation	1.00×	1.00×	1.00×	1.00×	1.02×	1.00×
CRCO	1.00×	1.00×	1.00×	1.00×	1.05×	1.00×
Effect-aware Sinkhorn-8	4.00×	4.00×	3.60×	1.00×	3.70×	1.00×

CRCO therefore requires no additional experience or model-FLOP allocation relative to paired GRPO, and its relative GPU time is $1.05\times$. Against ordinary unpaired GRPO, the corresponding multiplier is $1.05/0.55 = 1.91\times$, together with $2\times$ generated tokens and simulator calls. The cost table reports normalized ratios only. At deployment, CRCO keeps the unchanged single-branch policy interface.

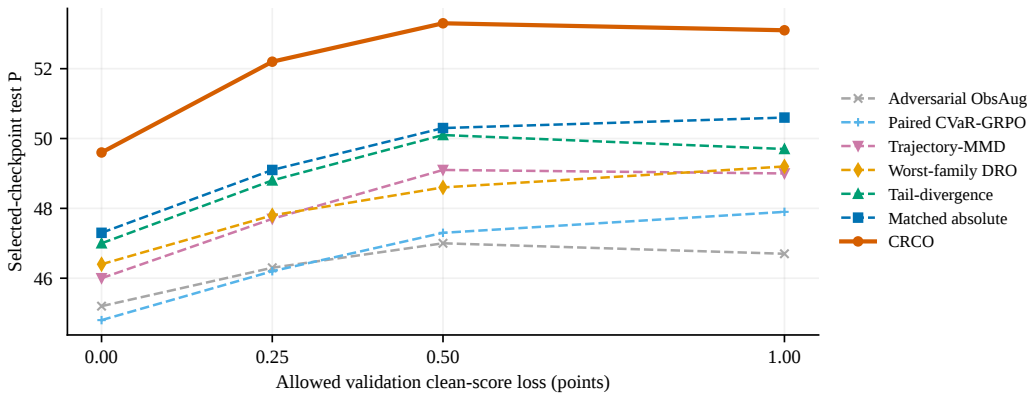


Figure 11: Test P at validation-selected clean-loss allowances; CRCO is highest at every evaluated allowance. Checkpoints are selected on validation tasks and scored on disjoint test tasks. Each marker is one reported aggregate, and segments connect the four evaluated allowances at their actual spacing. Curves include all seven objectives and preserve the observed decreases at the largest allowance.

M.5 Validation-selected regularization frontiers

To compare methods at equal clean sacrifice rather than at a single coefficient, each method's six-point coefficient grid $\{0, 0.25, 0.5, 1, 1.5, 2\}$ times its reference scale is evaluated on validation tasks. At clean-loss allowance δ_C, we select the checkpoint with the highest validation P among those within δ_C points of that method's best validation C and evaluate it once on disjoint test tasks. Because the constraint is method-specific, comparisons are aligned by attainable clean frontier, and all selection is nested inside the training seed. Figure 11 therefore reports robustness at equal within-method clean sacrifice across the complete coefficient grid.

Along the frontier, CRCO exceeds the strongest comparator at every allowance, by 2.3, 3.1, 3.0, and 2.5 points in order of increasing clean-loss allowance. At the prespecified 0.50 allowance, the paired margin is 3.0 points (95% interval $[1.7, 4.3]$). Independent test P is not constrained to be monotone in the validation allowance: four of 21 adjacent test transitions decline slightly after selection, including CRCO from 0.50 to 1.00, whereas the validation envelope is monotone by construction. Tail-divergence approaches matched absolute propagation on ordinary perturbations and improves upper-tail outcomes (Table 30), yet CRCO exceeds both along the validation-selected frontiers.

Regularization-strength sensitivity. Scaling both pair penalties jointly over $\{0, 0.25, 0.50, 1, 1.5, 2\}$ at a fixed ratio yields six-seed macro P scores of 41.9, 49.1, 52.0, 53.3, 52.7, and 51.5, respectively. The multiplier-0 endpoint is paired GRPO on the six-seed sweep panel (C=60.4, slope 1.14), which is distinct from both unpaired GRPO and the eight-seed main panel, and multiplier 1 reprints the selected six-seed reference. The validation-selected multiplier 1 is an interior optimum. Doubling it lowers the fitted slope further, from 0.69 to 0.59, but also lowers clean score from 61.6 to 60.1, entropy from 1.42 to 1.35, and semantic response from 83.5 to 80.4. Stronger regularization alone therefore does not explain the task-score advantage of CRCO.

M.6 Objective decomposition and depth

Table 37: 2×2 decomposition of the CRCO objective into its injection term and its affine contraction term, macro-averaged over four benchmarks; columns are distributional summaries. The injection term alone mainly lowers the zero-class leakage quantile, the affine contraction term alone mainly lowers the fitted slope, and the full objective gives the best value in every column.

Objective	$\widehat{q}^{\,\mathrm{zero}}_{0.95} \downarrow$	$\widehat{b}_1 \downarrow$	P $\uparrow$
Task only (paired GRPO)	0.040	1.13	42.0
$+\lambda_{\mathrm{inv}} \widetilde{D}_0$ only	0.014	1.10	47.3
+ affine contraction only	0.036	0.72	49.6
Full CRCO	**0.012**	**0.69**	**53.3**

Table 38: Architectural and degeneration controls, averaged over four benchmarks. Replacing the effect metric by text action distance, restricting pairs to a single step, or replacing the affine hinge by matched absolute propagation each lowers P relative to CRCO; removing task reward attains the lowest slope but sharply lowers C, P, entropy, and semantic response.

Variant	C ↑	P ↑	Entropy	Semantic response	Invalid (%)	$\widehat{b}_1 \downarrow$
CRCO ($K = 5$)	**61.6**	**53.3**	1.42	**83.5**	**7.9**	0.69
Text action distance	60.8	49.8	1.39	82.7	9.8	0.78
Single-step pairs ($K = 1$)	61.3	49.2	1.43	82.9	10.4	0.83
Affine hinge → matched absolute	60.0	50.7	1.42	82.8	9.3	0.81
No task reward	46.8	38.6	0.71	59.4	14.8	**0.61**

The matched-absolute ablation retains the injection term, coupling, and computation specified in Equation 17, so it differs from CRCO only in the propagation functional. Removing task reward lowers clean score by 14.8 points and approximately halves action entropy, demonstrating that task reward is needed to maintain capability and semantic responsiveness; the lowest slope in the table thus coincides with degenerate behavior. Increasing propagation depth improves P up to the selected horizon and then saturates: for $K \in \{1, 3, 5, 8\}$, P is 49.2, 52.0, 53.3, and 53.5, with mean slopes of 0.83, 0.74, 0.69, and 0.67. Every setting generates the same complete task-return continuation, and increasing K only scores additional edges from that continuation. Relative to $K = 5$, $K = 8$ therefore keeps generated tokens, simulator calls, and model FLOPs at $1.00\times$, while metric extraction and reward assembly rise to $1.46\times$ and $1.22\times$, respectively. Because these components occupy 17% and 6% of the $K = 5$ runtime, end-to-end runtime rises to $1.09\times$ (benchmark–seed range 1.05–$1.13\times$). The selected $K = 5$ therefore captures most of the benefit, and $K = 8$ adds only 0.2 P points.

N Backbone and Transfer Tests

Table 39: Transfer diagnostics on two additional backbones, with six seeds per method and scores macro-averaged over benchmarks. Matched absolute and constant hinge retain the CRCO injection term and the paired compute budget. On both backbones, CRCO attains the highest C and P and the lowest leakage quantile and adjacent-cost slope among the four methods. This table preserves the original focused diagnostic subset; the aligned six-objective comparison is reported in Table 1.

Backbone	Method	C	P	$\widehat{q}^{\,\mathrm{zero}}_{0.95}$	$\widehat{b}_1$
Qwen2.5-14B	GRPO	64.8	47.0	0.036	1.10
	Matched absolute	64.4	54.5	0.013	0.79
	Constant hinge	64.9	55.6	0.012	0.75
	CRCO	65.7	59.7	0.010	0.66
Llama-3.1-8B (Grattafiori et al., 2024)	GRPO	58.7	39.4	0.044	1.20
	Matched absolute	58.2	48.2	0.016	0.87
	Constant hinge	58.9	48.3	0.015	0.84
	CRCO	59.4	50.2	0.014	0.72

To evaluate cross-benchmark transfer, a rank-16 stability adapter trained on ALFWorld and ScienceWorld is merged into a fixed GRPO policy. On WebShop, P increases from 52.4 to 57.0, against 62.7 for in-domain CRCO; on τ-bench, P increases from 31.0 to 36.8, against 43.6 in domain. The source-trained adapter thus recovers approximately half of the in-domain improvement on both targets, indicating partial rather than complete transfer.

Deployment sensitivity. The primary training-to-deployment audit yields 0.94 checkpoint-rank agreement between $T = 0.7$ sampling and greedy decoding. An eight-seed deployment audit then holds semantic geometry fixed while applying weight quantization, rank truncation, or $T = 0.9$ sampling to frozen policies. With paired tasks and a disjoint 64-sample OT audit, the CRCO P advantage over the corresponding GRPO variant ranges from 9.1 to 11.8 points across these shifts. The refitted profiles characterize deployment sensitivity empirically, and Proposition 9 supplies the uniform bound under its kernel-displacement assumption.

O Recent Models and Broader Interaction Settings

Scope. This appendix extends the test of the state-conditioned propagation hypothesis along four axes: model generation, dense versus MoE architecture, interaction mechanism, and policy-generated occupancy. The five added backbones raise the evaluation to nine model variants, and the four added settings raise coverage to eight interactive settings; together they supply the transfer evidence for RQ1. The matched comparison consists of two complete blocks that are reported separately, a full 9×4 reference-prefix block and a full 9×4 reset-to-terminal block. Three backbones, Qwen3.5-9B, Qwen3.8-27B, and Gemma-4-31B-IT, additionally receive six-objective reset-to-terminal panels on the four added settings and complete reset-to-terminal evaluation on the four original settings.

O.1 Model identity, reasoning, and resource matching

The added models are Qwen3.5-9B, Qwen3.6-35B-A3B, Qwen3.8-27B, Gemma-4-26B-A4B-IT, and Gemma-4-31B-IT. The official release history dates the three Qwen releases to March 2, April 16, and August 14, 2026 (Qwen Team, 2026a;b;c), and both Gemma variants are listed in the official checkpoint registry (Google DeepMind, 2026). The mix of MoE and dense variants broadens architectural coverage, but differences between models also reflect their distinct pretraining and post-training histories. Every method comparison is therefore made within a single backbone.

The prescribed model identifiers are `Qwen/Qwen3.5-9B`, `Qwen/Qwen3.6-35B-A3B`, `Qwen/Qwen3.8-27B`, and the official Gemma registry entries `GEMMA4_26B_A4B_IT` and `GEMMA4_31B_IT`. All runs use text observations and each checkpoint's native chat template. Native reasoning is retained, and all reasoning and action tokens are charged to the same within-backbone budget. Context is capped at 16,384 tokens and each action generation at 1,024 tokens, and every objective follows identical retention and parsing rules. Reasoning text remains internal to the agent: only the parsed action changes simulator state. If the token cap is reached without a parseable action, the benchmark's invalid-action rule consumes the step. Truncation and invalid-action rates are part of the execution ledger.

The extension retains $K = 5$, $\alpha = 0.30$, $\bar{\rho} = 0.65$, four coupled pairs per group, training temperature 0.7, and greedy evaluation. LoRA rank and scale are $64/128$, with trainable updates on the attention projections and attention output projections; vision encoders and MoE routers remain frozen. The learning rate is 10^{-5} for the 9B model and 5×10^{-6} for the larger models, and it is held fixed across objectives within a backbone. The propagation-penalty coefficient is chosen from a common search over six values, $\{0.10, 0.20, 0.35, 0.50, 0.70, 1.00\}$, while the injection coefficient remains 0.15. Core benchmark tolerances follow Equation 16, and the added settings use $\bar{q} = 0.01$. The selection rule maximizes development P subject to losing at most one clean point relative to paired GRPO, and the constant and shuffled hinges are calibrated and selected by the same development procedure.

Each added comparison uses six training seeds. The tables report aggregate means, and uncertainty intervals appear only for analyses with seed-level inference. Every method receives the same per-backbone, per-benchmark ceilings of eight million generated tokens and 200,000 simulator transitions, and training stops as soon as either ceiling is reached. Hyperparameter search uses the same number of candidates and the same per-candidate ceiling for every method. Because equal ceilings do not imply identical realized use, the execution ledger records tokens, transitions, update batches, and FLOPs separately. HCAPO relies on hindsight-verification forwards rather than a separately trained value critic, and its verification tokens and additional forward FLOPs are charged to the corresponding resource totals. For MoE backbones, active parameter counts are kept distinct from total storage and measured wall-clock cost.

O.2 Added benchmark settings and semantic metrics

Each added setting introduces a different route from local misunderstanding to later state divergence. WebArena (Zhou et al., 2023) contributes persistent multi-site web state, and WorkArena++ L2 (Boisvert et al., 2024) contributes compositions of knowledge-work operations. The telecom domain of τ^2-bench (Barres et al., 2025) contributes dual control, in which both user and assistant actions change the environment, and τ-Knowledge (Shi et al., 2026) contributes retrieval-dependent conversational decisions.

WebArena is evaluated on its 812 public evaluation tasks, whereas its training/development tasks are separately authored on disjoint intent templates. WorkArena++ uses 240 study-specific held-out L2 instances, stratified by compositional task family, whose configuration seeds and referenced entities are disjoint from training and development; this count defines a study panel, not the official benchmark

size. Telecom and banking knowledge use the complete released base task lists in `tau2-bench v1.0.1`, and their published intents remain test-only. For each added setting, 1,000 training and 256 development scenarios are generated from public interfaces and policies, excluding held-out intent templates and expected solutions. The original τ-bench stays on its original 2024 release and is not merged with either new domain, whereas the banking-knowledge grading fixes in v1.0.1 are part of the specified setting.

Every added setting is scored by its native terminal success criterion on a 0–100 display, and no dense shaping reward substitutes for success in the corresponding tables. WebArena and WorkArena++ use text accessibility observations and native state validators. Telecom and knowledge use text interaction with a frozen Gemma-3-12B-IT user simulator at temperature 0.7, with one candidate per task per simulator trial and four simulator trials per task; pass^1 averages over trials rather than selecting the best one. The same user simulator serves all assistant backbones and objective arms, and its prompts, retained state, and random state are stored in every replay checkpoint. Knowledge retrieval uses the release's BM25-only configuration over a fixed public document collection, and admissible renderings leave indexed content, document identifiers, retrieval scores, and result ranking unchanged. The assistant receives neither hidden simulator predicates nor the expected answer.

Table 40: Semantic metrics and action limits specified for the four added settings. Block weights sum to one within each setting. Belief/history components use Equation 19, and all extractors and vocabularies are fixed from the training interfaces.

Setting	Semantic blocks (weight)	Action limit
WebArena	Goal-relevant database fields (0.35), page/form state (0.25), commitments and evidence (0.25), valid-action affordances (0.15)	30
WorkArena++ L2	Completed workflow predicates (0.35), record deltas (0.30), prerequisites and commitments (0.20), affordances (0.15)	50
τ^2 telecom	Network/device state (0.30), user constraints (0.25), completed workflow steps (0.25), irreversible operations (0.20)	40
τ-Knowledge	Account/task state (0.25), retained document evidence (0.30), user constraints and commitments (0.30), affordances (0.15)	40

Admissible transformations are defined per setting so that observation equivalence does not depend on the policy under evaluation. Web transformations reorder unordered metadata fields or change serialization whitespace while preserving node IDs, roles, labels, focus, and action mappings; they never permute a task-relevant ordered list or change the rendered page. Telecom transformations reorder independent tool-result fields without changing network state or the user's available actions. Knowledge transformations change the result-wrapper serialization while retaining all text, citations, scores, and document order. Any transformation that alters the retained tokenized history is ineligible. The existing paired-rollout method requires concrete environment access. For web environments, a replay comprises the database snapshot, session/authentication state, browser storage, URL, form state, and the action history needed to reconstruct the checkpoint; a copied DOM tree alone does not restore the transition process. A replay is eligible only when the native task predicates and action affordances match the checkpoint, and for dual-control tasks both assistant and user states must be restored.

O.3 Complete recent-model objective matrix

Under reference-prefix execution, all five recent backbones show macro P advantages of 0.5–5.0 points over matched absolute propagation after rounding. The size and task pattern of this benefit depend jointly on the backbone and its recovery opportunity rather than on a fixed benchmark template. The advantage is not uniform at the task level: Gemma-4-26B retains a negative ScienceWorld cell despite its positive macro gain, and task-level signs vary within four of the five recent-backbone blocks. The ordering among matched absolute, constant, and shuffled hinges also changes across models. Higher base capability raises the task-score level, but it neither adds a fixed robustness increment nor orders objective gains monotonically by model size.

Table 41: Qwen3.5-9B: clean/perturbed (C/P) scores of each objective under reference-prefix execution, a separate estimand from the other execution protocol. Macro gives equal weight to the four displayed settings; the highlighted row is CRCO.

Objective	ALFWorld	ScienceWorld	WebShop	τ-bench	Macro
Paired GRPO	82.1 / 68.3	59.6 / 41.7	76.8 / 61.2	58.3 / 40.8	69.2 / 53.0
One-step consistency	81.7 / 73.6	59.9 / 47.9	76.0 / 66.8	58.6 / 46.5	69.1 / 58.7
Matched absolute	82.6 / 79.5	60.2 / 51.6	76.5 / 69.3	59.0 / 50.8	69.6 / 62.8
Constant hinge	82.8 / 76.9	60.7 / 52.5	76.7 / 69.5	58.7 / 51.5	69.7 / 62.6
Shuffled-current hinge	82.2 / 76.6	60.4 / 52.1	76.8 / 69.4	59.2 / 51.2	69.7 / 62.3
CRCO	**83.2 / 79.3**	**61.9 / 58.7**	**77.1 / 74.1**	**60.1 / 58.3**	**70.6 / 67.6**

Table 42: Qwen3.6-35B-A3B: clean/perturbed (C/P) scores of each objective under reference-prefix execution, a separate estimand from the other execution protocol. Macro gives equal weight to the four displayed settings; the highlighted row is CRCO.

Objective	ALFWorld	ScienceWorld	WebShop	τ-bench	Macro
Paired GRPO	85.4 / 73.8	65.8 / 49.4	79.6 / 66.5	66.7 / 48.9	74.4 / 59.7
One-step consistency	85.1 / 78.6	65.3 / 54.5	79.8 / 71.3	66.2 / 55.1	74.1 / 64.9
Matched absolute	85.9 / 84.0	66.0 / 57.8	79.3 / 75.4	67.1 / 62.4	74.6 / 69.9
Constant hinge	86.1 / 81.0	66.2 / 58.7	79.9 / 73.8	67.3 / 59.4	74.9 / 68.2
Shuffled-current hinge	85.6 / 80.7	66.4 / 58.4	79.5 / 73.6	66.8 / 59.1	74.6 / 68.0
CRCO	**86.7 / 83.4**	**67.4 / 63.6**	**80.2 / 75.4**	**68.0 / 63.2**	**75.6 / 71.4**

Table 43: Qwen3.8-27B: clean/perturbed (C/P) scores of each objective under reference-prefix execution, a separate estimand from the other execution protocol. Macro gives equal weight to the four displayed settings; the highlighted row is CRCO.

Objective	ALFWorld	ScienceWorld	WebShop	τ-bench	Macro
Paired GRPO	89.3 / 81.2	75.6 / 68.0	83.0 / 72.9	75.1 / 69.9	80.8 / 73.0
One-step consistency	89.7 / 85.0	75.2 / 62.6	82.8 / 76.0	75.4 / 66.0	80.8 / 72.4
Matched absolute	90.1 / 87.5	75.8 / 66.8	83.3 / 78.9	75.6 / 67.6	81.2 / 75.2
Constant hinge	90.0 / 87.1	76.0 / 66.5	83.1 / 79.4	75.9 / 69.1	81.3 / 75.5
Shuffled-current hinge	89.8 / 87.0	75.7 / 69.0	83.4 / 79.5	75.5 / 66.9	81.1 / 75.6
CRCO	**90.5 / 88.1**	**76.3 / 74.3**	**83.5 / 83.0**	**76.4 / 75.4**	**81.7 / 80.2**

Table 44: Gemma-4-26B-A4B-IT: clean/perturbed (C/P) scores of each objective under reference-prefix execution, a separate estimand from the other execution protocol. Macro gives equal weight to the four displayed settings; the highlighted row is CRCO.

Objective	ALFWorld	ScienceWorld	WebShop	τ-bench	Macro
Paired GRPO	83.6 / 70.2	62.9 / 44.8	77.9 / 63.0	61.4 / 42.6	71.5 / 55.2
One-step consistency	83.4 / 74.8	63.2 / 50.2	77.6 / 68.1	61.9 / 48.3	71.5 / 60.4
Matched absolute	84.0 / 82.6	63.6 / 59.6	78.2 / 73.0	62.5 / 56.8	72.1 / 68.0
Constant hinge	84.2 / 78.3	63.5 / 54.6	78.0 / 70.6	62.2 / 52.8	72.0 / 64.1
Shuffled-current hinge	83.8 / 77.8	63.9 / 54.0	78.3 / 70.4	62.4 / 52.6	72.1 / 63.7
CRCO	**84.9 / 83.1**	**65.0 / 59.1**	**78.7 / 74.1**	**63.7 / 57.6**	**73.1 / 68.5**

Table 45: Gemma-4-31B-IT: clean/perturbed (C/P) scores of each objective under reference-prefix execution, a separate estimand from the other execution protocol. Macro gives equal weight to the four displayed settings; the highlighted row is CRCO.

Objective	ALFWorld	ScienceWorld	WebShop	τ-bench	Macro
Paired GRPO	88.4 / 78.9	72.3 / 56.7	82.1 / 70.6	72.5 / 55.2	78.8 / 65.4
One-step consistency	88.0 / 81.8	72.8 / 62.9	81.9 / 73.8	72.1 / 59.4	78.7 / 69.5
Matched absolute	88.7 / 84.3	73.2 / 64.7	82.5 / 79.1	72.8 / 65.0	79.3 / 73.3
Constant hinge	89.0 / 85.4	73.0 / 64.1	82.4 / 75.6	73.1 / 67.1	79.4 / 73.1
Shuffled-current hinge	88.6 / 83.2	73.4 / 68.8	82.3 / 75.1	72.7 / 66.8	79.3 / 73.5
CRCO	**89.5 / 89.1**	**74.2 / 69.9**	**82.9 / 78.8**	**74.0 / 71.6**	**80.2 / 77.4**

O.4 Complete extension matrix and baseline roles

Table 46: Qwen3.5-9B: clean/perturbed (C/P) scores of each objective under reset-to-terminal execution, a separate estimand from the other execution protocol. Macro gives equal weight to the four displayed settings; the highlighted row is CRCO.

Objective	WebArena	WorkArena++	τ^2 telecom	τ-Knowledge	Macro
Paired GRPO	31.4 / 19.3	18.6 / 12.1	68.7 / 48.9	14.8 / 10.4	33.4 / 22.7
HCAPO	34.6 / 21.6	22.3 / 14.9	71.4 / 51.6	17.5 / 12.6	36.5 / 25.2
One-step consistency	31.2 / 23.0	18.9 / 13.8	68.2 / 53.7	14.6 / 11.7	33.2 / 25.6
Matched absolute	32.3 / 26.5	20.1 / 19.9	69.6 / 58.4	16.1 / 13.2	34.5 / 29.5
Constant hinge	32.8 / 26.4	20.5 / 16.5	70.1 / 59.2	16.4 / 13.7	35.0 / 29.0
CRCO	**34.1 / 29.6**	**22.0 / 18.7**	**72.0 / 64.6**	**17.8 / 15.9**	**36.5 / 32.2**

Table 47: Qwen3.8-27B: clean/perturbed (C/P) scores of each objective under reset-to-terminal execution, a separate estimand from the other execution protocol. Macro gives equal weight to the four displayed settings; the highlighted row is CRCO.

Objective	WebArena	WorkArena++	τ^2 telecom	τ-Knowledge	Macro
Paired GRPO	55.8 / 41.2	32.4 / 23.7	74.3 / 55.2	25.3 / 17.9	47.0 / 34.5
HCAPO	59.0 / 44.8	35.8 / 25.1	76.8 / 58.7	29.1 / 22.7	50.2 / 37.8
One-step consistency	56.1 / 46.3	32.1 / 24.0	74.0 / 61.3	25.1 / 20.9	46.8 / 38.1
Matched absolute	56.9 / 48.1	33.6 / 29.2	75.2 / 66.8	26.7 / 22.6	48.1 / 41.7
Constant hinge	57.4 / 49.5	34.0 / 27.6	75.6 / 65.9	27.3 / 23.8	48.6 / 41.7
CRCO	**58.7 / 55.1**	**35.1 / 30.6**	**77.4 / 71.2**	**28.4 / 27.3**	**49.9 / 46.1**

Table 48: Gemma-4-31B-IT: clean/perturbed (C/P) scores of each objective under reset-to-terminal execution, a separate estimand from the other execution protocol. Macro gives equal weight to the four displayed settings; the highlighted row is CRCO.

Objective	WebArena	WorkArena++	τ^2 telecom	τ-Knowledge	Macro
Paired GRPO	50.7 / 33.7	30.1 / 22.3	70.1 / 49.2	23.7 / 17.8	43.7 / 30.8
HCAPO	53.9 / 37.9	33.5 / 24.9	72.4 / 52.3	27.4 / 19.9	46.8 / 33.8
One-step consistency	51.0 / 38.6	30.7 / 24.6	70.7 / 55.5	24.2 / 20.3	44.2 / 34.8
Matched absolute	51.8 / 45.0	31.8 / 25.8	71.0 / 60.9	25.0 / 24.9	44.9 / 39.2
Constant hinge	52.2 / 42.6	32.0 / 27.2	71.8 / 60.3	25.6 / 22.7	45.4 / 38.2
CRCO	**53.4 / 48.3**	**33.2 / 30.9**	**73.2 / 65.1**	**26.9 / 25.2**	**46.7 / 42.4**

These tables expand the six-objective reset-to-terminal panels for the three designated recent backbones. Table 2 reports the reset-to-terminal comparison between matched absolute propagation and CRCO for all nine backbones; the remaining six backbones therefore enter the 72-cell matrix without being counted as six-objective replications.

Each baseline isolates one design element. Paired GRPO isolates the effect of adding robustness penalties at the same branch count, and one-step consistency isolates injection control. Matched absolute propagation retains the injection term and all paired observations, which makes it the closest

comparator for conditional propagation; the constant hinge additionally matches selective activation. HCAPO (Tan et al., 2026) tests whether stronger long-horizon credit assignment closes the robustness gap, and the shuffled-current intervention isolates state correspondence. Every learned baseline shares the assistant, tools, prompts, benchmark reward, and within-backbone resource ceiling. External leaderboard results serve only as capability anchors.

On the four added settings, CRCO exceeds matched absolute propagation in macro P by 2.7, 4.4, and 3.2 points for Qwen3.5-9B, Qwen3.8-27B, and Gemma-4-31B-IT, respectively. The per-setting effect is heterogeneous: Qwen3.5 has one local loss, whereas Qwen3.8 and Gemma-4-31B remain positive on all four settings, with their largest gains on telecom, WebArena, and WorkArena++, respectively. This pattern supports the recovery hypothesis without requiring a uniform per-task effect. Appendix Q relates the base score levels to public anchors and states where their source protocols differ from ours.

O.5 From reference prefixes to complete task execution

Table 49: Reset-to-terminal results of three recent backbones on the four original benchmark settings. Each candidate generates its own prefix from task reset. A is matched absolute propagation; ΔP is CRCO minus A. CRCO improves P in all but two cells, both small WebShop losses. These rows define an estimand separate from the reference-prefix results in Table 1 and are not pooled with them.

Backbone	Benchmark	A C/P	CRCO C/P	ΔP
Qwen3.5-9B	ALFWorld	80.6 / 75.0	81.3 / 79.6	+4.6
	ScienceWorld	56.8 / 50.1	58.0 / 52.3	+2.2
	WebShop	74.9 / 69.3	75.4 / 69.0	-0.3
	τ-bench	56.1 / 48.2	57.5 / 54.0	+5.8
Qwen3.8-27B	ALFWorld	88.9 / 85.7	89.6 / 86.8	+1.1
	ScienceWorld	73.1 / 64.4	74.0 / 70.3	+5.9
	WebShop	81.7 / 77.0	82.1 / 80.0	+3.0
	τ-bench	72.8 / 66.4	74.0 / 69.0	+2.6
Gemma-4-31B-IT	ALFWorld	87.0 / 82.9	88.1 / 86.6	+3.7
	ScienceWorld	70.4 / 62.5	71.7 / 65.9	+3.4
	WebShop	80.8 / 77.3	81.4 / 77.2	-0.1
	τ-bench	69.7 / 61.6	71.2 / 67.7	+6.1

In this protocol, each assistant generates its own history from task reset. For every task and replicate, an intervention step is drawn uniformly from the first 60% of the benchmark action limit before execution begins. A task that terminates before this step keeps its terminal outcome in the score denominator, and an ineligible observation receives the identity transformation. The clean and perturbed executions share the candidate's own prefix up to the intervention and branch afterward. Task IDs, scheduled intervention steps, and simulator seeds are shared across objectives, but prefixes may differ. This design avoids conditioning the comparison on method-specific survival or on the state visitation of a reference policy.

On the four original core benchmarks, the reset-to-terminal advantages of CRCO over matched absolute propagation are 3.1, 3.2, and 3.3 points for the three backbones, with two slightly negative WebShop cells; these advantages are computed separately from the reference-prefix contrasts and are not pooled with them. Reset-to-terminal evaluation measures the policy's own occupancy, whereas the common-prefix matrix identifies the local objective mechanism under aligned states, so the two protocols answer distinct questions. The four added benchmark settings use reset-to-terminal execution throughout. Deployment still uses a single assistant rollout and adds no paired probes.

Aggregation and arithmetic. Base capability and task difficulty determine the score level, whereas horizon, recoverability, and early action commitment determine the robustness margin. The six-objective and three-backbone end-to-end appendix tables are regenerated from 216 result records, and the 72-cell matched summary is regenerated from 144 matched-comparison records; policy rows shared across views are kept numerically identical. All aggregation uses decimal arithmetic and a single round-half-up display rule.

P Hyperparameters and Statistical Inference

Benchmark versions and score conventions. We use the original 2024 airline/retail release of τ-bench corresponding to Yao et al. (2024). ScienceWorld normalized score and WebShop reward are

displayed on a 0–100 scale, and ALFWorld success and τ-bench pass[1] are percentages. Following the original protocol, the τ-bench score first averages trials within each task and tasks within each domain, and then weights the retail and airline domains equally.

The published values below follow the protocols of their source papers and provide external context only; Table 32 reports results under our study protocol. The original ScienceWorld DRRN average is 0.17 (17.0 on our display). The original WebShop IL+RL and human rewards are 62.4 and 82.1, with success rates of 28.7% and 59.6%. For Qwen2.5-7B, Tan et al. (2026) report ALFWorld success of 77.6 for GRPO and 91.4 for HCAPO, and WebShop rewards of 79.3 and 85.1, respectively (WebShop success: 66.1% and 73.8%). The published HCAPO ALFWorld value is the variant without temporal smoothing; the smoothed variant reports 96.9, and the two variants are not interchangeable. The original τ-bench Table 2 reports GPT-4o pass[1] of 61.2/35.2 on retail/airline, a domain-macro average of 48.2. The Gemma 3 Technical Report gives Gemma-3-12B-IT zero-shot scores of 60.6 on MMLU-Pro, 40.9 on GPQA Diamond, 83.8 on MATH, and 69.5 on Global MMLU-Lite; these static scores establish model-scale context but are not direct anchors for the interactive C/P estimand.

Table 50: Optimization and evaluation settings used throughout the experiments.

Backbone	Qwen2.5-7B	LoRA rank / α	64 / 128
Learning rate	10^{-5}	KL coefficient	0.02
Group size	8 branches	Context	16,384
Pair depth K	5	$\bar{\rho}$ / $\bar{q}$	0.65 / 0.009, 0.011, 0.010, 0.012
$(\lambda_{\mathrm{inv}}, \lambda_{\mathrm{ctr}})$	(0.15, 0.35)	State weight α	0.30
Seeds	6 broad / 8 confirmatory	Train / evaluation sampling	$T = 0.7$ / greedy

Training proceeds in three stages: a 50-task debug stage ($K = 2$, one seed, at most 200 updates), a one-seed 10% development screen at $K = 5$, and full runs. The environment image, model revision, prompts, transformation manifest, task IDs, and statistical criteria are frozen after the development screen. The screen requires at least 10% lower divergence AUC and a 2-point paired-retention gain over both observation augmentation and matched absolute propagation, with no more than 1.5 clean points lost. Development uses training tasks only, and debug and screening runs are excluded from the inferential sample.

Two estimands are distinguished. The fixed-policy estimand averages paired task differences conditional on a trained checkpoint, whereas the training estimand additionally averages over independent training seeds. Whenever a table reports mean $\pm$ standard deviation over seeds, the standard deviation is the Bessel-corrected sample standard deviation of unrounded seed-level scores, and rounding occurs only after aggregation. Main C/P contrasts use the validation-selected checkpoint of each run, whereas the confirmatory retention analysis averages the prespecified checkpoint schedule within each seed; its statistic is the equally weighted mean of four benchmark-level $100P_b/C_b$ contrasts. A 10,000-replicate bootstrap resamples the eight matched seed blocks together with their complete method–benchmark–checkpoint panels and then resamples paired tasks within each selected block. The same task resample is reused wherever that task recurs across checkpoints, which preserves checkpoint dependence and the fixed evaluation schedule. Seed counts differ by panel: the conventional-baseline panel in Table 32 uses the first six Qwen2.5-7B seeds, the matched Qwen2.5-7B anchor rows in Tables 1 and 2 use all eight, and Gemma-3-12B and every other backbone use six. The eight-seed mechanism comparisons cover paired GRPO, one-step consistency, matched absolute, both hinge controls, direct-state training, CRCO, and observation augmentation. Direct-state and competing-exit contrasts provide focused mechanism analyses.

Values that recur across the mechanism, ablation, cost, and preprocessing tables are labeled reference rows drawn from these shared policy panels. They make within-table contrasts readable and are not counted as independent replications: a result is treated as independent evidence only when the text specifies a new training run, seed block, state-pair panel, or intervention assignment.

The confirmatory analysis applies three procedures to the same per-seed checkpoint averages: a hierarchical bootstrap, a wild-cluster bootstrap, and a paired-seed sign-flip test. The wild bootstrap uses seed-level Rademacher weights, and the sign-flip test enumerates all 2^8 assignments for the studentized contrast.

Tail risk and checkpoint equivalence. Terminal quantiles, empirical CVaR, and irreversible-failure rates are first computed across held-out tasks within each fixed policy unit. The confirmatory contrast between CRCO and matched absolute propagation uses eight paired seeds, whereas com-

Table 51: Eight-seed retention effects of CRCO, averaged over the fixed checkpoint schedule, with seed-block–task bootstrap intervals. The benchmark-stratified aggregate weights all four benchmarks equally. The contrast with matched absolute propagation is the confirmatory test: it is positive on every benchmark, and its interval excludes zero in aggregate and on ALFWorld and ScienceWorld but not on the remaining two benchmarks. All contrasts with observation augmentation (ObsAug) exclude zero.

Benchmark	Δ vs. matched absolute	Δ vs. ObsAug
ALFWorld	$+2.0$ $[0.4, 3.5]$	$+4.6$ $[2.8, 6.3]$
ScienceWorld	$+3.1$ $[0.8, 5.2]$	$+10.3$ $[7.5, 12.8]$
WebShop	$+1.3$ $[-0.2, 2.7]$	$+7.4$ $[5.5, 9.1]$
τ-bench	$+1.8$ $[0.0, 3.6]$	$+14.1$ $[11.0, 16.8]$
Benchmark-stratified	$\mathbf{+2.0}$ $[\mathbf{0.7}, \mathbf{3.2}]$	$\mathbf{+9.1}$ $[\mathbf{7.4}, \mathbf{10.8}]$

parisons with the two tail-specific baselines use six. Tail-recurrence coverage is evaluated over benchmark $\times$ training seed $\times$ three checkpoints fixed at 50%, 75%, and 100% of updates, which gives $4 \times 8 \times 3 = 96$ units for the four eight-seed rows and $4 \times 6 \times 3 = 72$ for the two six-seed tail baselines; in Table 30 order, the coverage counts are 88/96, 91/96, 66/72, 89/96, 69/72, and 90/96. Because checkpoints are repeated measurements within a run, every contrast first aggregates them within benchmark–seed and then resamples seed as the outer unit, resampling tasks only inside a sampled checkpoint policy. Checkpoint selection is evaluated on the eight paired CRCO seeds against a prespecified ± 1 P-point equivalence margin: equivalence to full-horizon selection is declared only if the entire two-sided 95% interval for the paired difference lies inside that margin. Test tasks are opened once, after every selector has frozen its checkpoint.

Matched-gain inference. The matched-gain analysis treats seed as the outer unit and the matched checkpoint pair as an inner unit, resampling tasks only inside a policy. Balance, slope, and long-horizon outcomes are estimated on three disjoint task partitions, and the within-family analysis performs an independent matching procedure within each objective family.

Target-sweep inference. The unit of the randomized sweep is one separately optimized policy, and target effects are identified by within-block assignment among the five exchangeable rollout streams. The estimand averages paired target contrasts over 16 benchmark–initialization blocks. Exact randomization permutes labels only within a block, whereas the reported interval resamples blocks within benchmark. Probe tasks estimate the summaries each policy achieves, while the optimized policy remains the unit of treatment assignment.

For the associational mechanism model, p_{mbsc} denotes the probability of failing the benchmark's binary goal-completion criterion at a specified horizon stratum. It differs from score retention P/C, which need not be a probability. We fit

$$\begin{aligned}\operatorname{logit}(p_{mbsc}) = \beta_0 + \beta_1 \widehat{b}_{1,mbsc} + \beta_2 \widehat{q}^{\,\mathrm{zero}}_{0.95,mbsc} + \beta_3 \widetilde{D}_{0,mbsc} + \beta_4 H_{mbsc} \\ + \beta_5 C_{mbsc} + \beta_6 I_{mbsc} + u_b + v_{f(m)} \widehat{b}_{1,mbsc},\end{aligned} \tag{78}$$

for method m, benchmark b, seed s, and checkpoint c, with benchmark intercept u_b and method-family random slope $v_{f(m)}$. Here C is clean score, I is invalid-action rate, and H is remaining horizon. Binary task outcomes define the likelihood. AUROC, Brier, and calibration intervals resample complete seed panels because all tasks within a policy unit share the same fitted profile. Irreversible-failure analyses use their separately stated event definition, with terminal states treated as absorbing.

Sensitivity and power. The power sensitivity analysis uses eight independently initialized policies for each confirmatory method, paired task outcomes, benchmark strata, an assumed between-seed SD of 1.1–1.5 points, and task discordance of 0.07–0.10. The task ledger contains 1,099 distinct clusters across four benchmark-specific policies per seed, and repeated user or simulator draws retain their original task cluster. Under these variance assumptions, the minimum detectable effect at 80% power is 1.5–1.8 points. Confirmatory inference relies on independent training seeds and paired tasks, whereas the broad comparison matrix uses six seeds.

Q External Score Anchors and Protocol Calibration

This appendix separates model capability from the effects of adapters, task versions, prefixes, reasoning limits, and user simulators by placing published scores alongside, but never inside, our

measurements. The source values below were checked on September 23, 2026 against the original papers and official model cards, with model cards retrieved through a Hugging Face mirror where direct access failed. Every published score retains its original protocol and is not a CRCO measurement.

Table 52: Published external anchors with their original metrics and protocols, shown on a 0–100 display. Every value is reported by the cited authors under the stated source setting and is not a measurement from this study; the last column states the setting and the limits of its comparability with ours.

Source model / method	Score	Setting and comparability
Qwen2.5-7B GRPO	77.6 / 79.3	ALFWorld success / WebShop reward; HCAPO source protocol (Tan et al., 2026).
Qwen2.5-7B HCAPO	91.4 / 85.1	Same metrics; without temporal smoothing. The smoothed ALFWorld variant is 96.9.
Qwen3.5-9B	79.1	Model-card TAU2-Bench aggregate (airline fixes); not telecom-only (Qwen Team, 2026a).
Gemma-4-31B / 26B-A4B	76.9 / 68.2	Model-card Tau2 average over three domains (Google DeepMind, 2026).
Qwen3.8-27B	64.8	WebArena-Verified with OSWorld scaffold; distinct from original 812-task WebArena (Qwen Team, 2026c).
GPT-4o / GPT-4o-V	3.0 / 3.8	Original WorkArena++ L2, 235-instance agent curriculum; not L1 (42.7 / 41.8) (Boisvert et al., 2024).
GPT-5.2 high reasoning	24.48	Original τ-Knowledge BM25; terminal retrieval gives 25.52 (Shi et al., 2026).
Claude-4.5-Opus high	39.69	τ-Knowledge Gold documents; removes retrieval and is not a BM25 baseline (Shi et al., 2026).

Same-backbone comparison. Under our protocol, the Qwen2.5-7B HCAPO baseline reaches clean scores of 89.6/82.3 on ALFWorld/WebShop, compared with 91.4/85.1 under the source protocol, and scores 68.0/63.6 under our perturbation protocol. Our comparison therefore retains a strong credit-assignment baseline. State conditioning is identified by the matched-objective comparison, not by a requirement that CRCO win every broad-baseline cell.

Across objectives, Qwen3.5-9B telecom clean scores range from 68.2 to 72.0, below the 79.1 model-card aggregate. Our setting differs through the 1,024-token action cap, the greedy assistant, a separate simulator, and a single domain, and the cap does not reproduce the model card's reasoning configuration. Gemma-4-31B telecom clean scores of 70.1–73.2 remain close in scale to its 76.9 three-domain context value. Neither comparison constitutes a telecom leaderboard replication.

Difficulty and release calibration. WorkArena++ L2 is substantially harder than L1. Across the three recent models and six objectives, clean scores span 18.6–35.8 under targeted adapter learning on the specified training scenarios. The 2024 GPT-4o result is therefore a difficulty anchor, not a ceiling on later adapted models. Likewise, the newer WebArena-Verified score of 64.8 provides context but is not substituted for original-task success.

For banking knowledge, the original paper reports BM25 scores of 13.66–24.48 for its high-reasoning frontier agents, and even Gold-document access reaches at most 39.69. Under our adapter and reasoning budget, clean BM25 scores span 14.6–29.1, and CRCO perturbed scores are 15.9/27.3/25.2. These values are not directly comparable with the published ones: the paper's original grading differs from our v1.0.1 grading, the benchmark maintainers explicitly prohibit treating pre-fix scores as directly comparable, and the user simulator also differs. The anchors thus establish a difficulty scale, not a conversion formula between versions.

Interpretation of external anchors. ALFWorld success is kept separate from WebShop reward, and the original τ-bench, τ^2 telecom, and banking knowledge are treated as separate settings. Reference-prefix and reset-to-terminal scores are never pooled. Because no published result matches the model, adapter, and budget for every cell of the extension, the external anchors establish provenance and protocol comparability rather than serving as direct replications.

R Related Work

This appendix positions the two methodological contributions against the closest lines of work. For each line we state what it mainly targets and the specific respect in which our measurement

(Contribution 1, Challenge 1) or CRCO (Contribution 2, Challenge 2) differs; the matched controls of Section 5 test the optimization-side distinctions directly.

State abstraction and bisimulation. Lumpability and homomorphisms preserve aggregated dynamics (Buchholz, 1994; Givan et al., 2003); bisimulation metrics quantify approximate equivalence and can shape learned representations (Ferns et al., 2011; Zhang et al., 2021). *These lines mainly derive an equivalence from the dynamics of a given process*, so the abstraction is defined by the behavior it summarizes. *In contrast*, we fix task semantics before policy optimization and independently of the policy. A policy that responds differently to semantically equivalent states therefore does not redefine the abstraction; it incurs measurable quotient leakage, which the same geometry separates from residual amplification on positive-distance pairs (Challenge 1).

Markov-kernel and incremental stability. Contraction coefficients, coarse Ricci curvature, and Wasserstein perturbation theory characterize the stability of Markov kernels (Ollivier, 2009; Rudolf & Schweizer, 2018); contraction analysis and incremental input-to-state stability compare trajectories under distinct disturbances (Lohmiller & Slotine, 1998; Angeli, 2002; 2009). Lyapunov-based reinforcement learning constrains cumulative safety along individual trajectories (Chow et al., 2018). *These tools mainly analyze a given kernel or constrain a single trajectory. In contrast*, we integrate them with a fixed task-semantic quotient that a language policy may fail to preserve, and we model explicitly the resulting leakage, the realizability of the affine bound through the native agent–environment factorization, and first-exit consequences. CRCO also differs from Lyapunov constraints in its object: it controls the propagation of *paired* task-state deviations rather than a cumulative cost along one trajectory, and it converts the resulting profile into a training objective tied to executable action effects.

Agent recovery and credit assignment. Reflection and tool-grounded correction recover after detected failures (Yao et al., 2022b; Madaan et al., 2023; Shinn et al., 2023; Gou et al., 2023; Sun et al., 2023), and Agent-R trains such reflection through iterative self-training (Yuan et al., 2025). Stepwise rollback and AgentRewind restore earlier execution states (Li et al., 2025; Zhuang et al., 2026), while error-localized policy optimization, hindsight credit assignment, milestone-guided learning, and reinforcement-learning training frameworks for agents improve or reorganize delayed credit (Liang et al., 2026; Tan et al., 2026; Wang et al., 2026; Luo et al., 2025). *These methods mainly intervene after a failure is detected or when outcome reward is assigned. In contrast*, CRCO changes the training objective that governs subsequent separation: each step's threshold is indexed by the current state pair, so a successor distance is penalized only when it exceeds the allowance implied by the deviation it follows (Challenge 2). Penalties that depend only on distance marginals cannot express this distinction (Proposition 1), and deployment keeps a single-rollout interface without detection, rollback, or generation-time selection. The matched absolute, constant-threshold, and shuffled-current controls test this distinction by isolating successor suppression, activation, and state correspondence.

Robust objectives. Robust MDPs optimize over transition uncertainty (Iyengar, 2005; Nilim & El Ghaoui, 2005), while input-augmentation regularization provides a related way to expose a policy to variation (Kostrikov et al., 2020). Our observation-augmentation baseline adapts that idea to the benchmark renderers and does not claim to reproduce the cited pixel-based method. *These approaches mainly bound the worst case over a transition set or constrain sensitivity to the input itself. In contrast*, CRCO acts on the closed-loop response after an input variation has already induced a deviation, through state-conditioned paired propagation control; the two are therefore complementary. Appendix M compares CRCO with this study-specific observation-augmentation baseline and with resource-matched tail-divergence, worst-family DRO, and trajectory-MMD objectives.

Interactive benchmarks and backbones. WebArena, WorkArena++, τ^2-bench, and τ-Knowledge add persistent web state, compositional work, dual control, and retrieval (Zhou et al., 2023; Boisvert et al., 2024; Barres et al., 2025; Shi et al., 2026), and BrowserGym and WorkArena provide richer browser infrastructure (De Chezelles et al., 2024; Drouin et al., 2024); recent Qwen and Gemma families add dense and MoE backbones (Qwen Team, 2026a;b;c; Google DeepMind, 2026). *These benchmarks primarily score task success under their native observation rendering. In contrast*, we use them to measure degradation under admissible, semantics-preserving transformations; the WebArena and WorkArena++ protocols in Appendix O specify the replayable state and task-effect measurements that extend the four original settings. Within-backbone resource matching isolates the objective, and cross-backbone results test transfer.