EQUIVALENT REWARDS, UNEQUAL INTERFACES: REWARD COORDINATES IN HISTORY-CONDITIONED RL

Anonymous authors
Paper under double-blind review

ABSTRACT

History-conditioned reinforcement learning agents use rewards both to define an objective and as inputs for online adaptation. Does preserving the objective also preserve adaptation? We distinguish objective, information, and policy-interface equivalence, and study their separation using interventions that alter only the rewards visible to the actor. Across five independently trained AMAGO policies in each of two continuous-control task families, positive-affine reward transformations alter actions and original-task returns. HalfCheetah exhibits a non-monotone, training-coordinate-specific response; Ant-Fwd-Back reproduces closed-loop sensitivity and a direct, reversible reward-to-action pathway. Teacher-forced replay localizes effects before environmental feedback, while HalfCheetah activation patching identifies structured, action-relevant changes not captured by the decoded task axis. Recurrent RL^2 agents also exhibit strong coordinate sensitivity, whereas Decision Transformer controls show weaker sensitivity consistent with their weaker reliance on trajectory-specific reward evidence. Zero and donor interventions further connect coordinate sensitivity to reward reliance. These results demonstrate reward-interface dependence in learned adaptation: equivalent objectives and information need not produce equivalent behavior when reward is part of the policy interface.

1 INTRODUCTION

Reward functions specify what an agent should optimize, but their numerical representation is often treated as secondary. Positive scaling preserves preferences over expected return, and potential-based shaping preserves optimal policies under standard assumptions (Ng et al., 1999; Wiewiora, 2003). Work on reward identifiability, normalization, and scale-robust optimization likewise studies objectives, learned rewards, or training dynamics (van Hasselt et al., 2016; Schaul et al., 2021; Skalse et al., 2023; Sullivan et al., 2023; Wang et al., 2024). These guarantees do not determine the evaluation-time behavior of a fixed sequence policy that receives reward as an input.

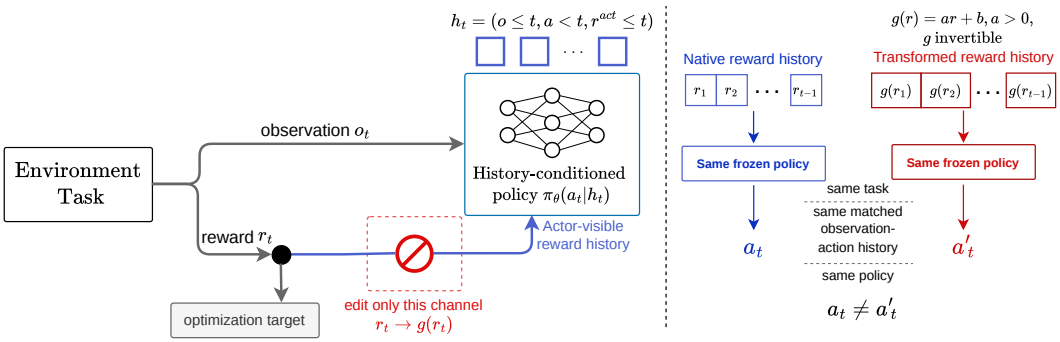


Figure 1: **Equivalent rewards need not be equivalent policy interfaces.** We transform only actor-visible reward while preserving the environment, original evaluation objective, and frozen policy. Even an invertible coordinate change can induce a different action on a matched observation–action history.

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.

For a history-conditioned agent, reward has two roles. It defines the learning objective and later becomes evidence used for online adaptation. Recurrent agents, task-inference methods, and attention-based in-context policies all condition action selection on histories of observations, actions, and rewards (Duan et al., 2016; Wang et al., 2017; Mishra et al., 2018; Rakelly et al., 2019; Zintgraf et al., 2020; Melo, 2022; Laskin et al., 2023; Grigsby et al., 2024a). The learned policy can attach task-specific meaning to a reward magnitude, offset, or temporal pattern even when another coordinate preserves both the objective and the information in the reward sequence.

This issue can arise without changing the physical task. Reward wrappers, multi-task pipelines, normalization conventions, and learned reward models can alter the coordinate presented to the actor while leaving the intended objective intact. Native-task competence does not specify how the learned computation should extend to such equivalent inputs, making the reward coordinate a structured source of underspecification (D'Amour et al., 2022). Unlike standard generalization shifts, the environment and task instance remain fixed (Cobbe et al., 2019; Song et al., 2019).

We study this gap with a matched causal question: with the task, evaluation objective, policy weights, and physical history fixed, does an invertible change of actor-visible reward alter the next decision? We distinguish *objective equivalence*, *information equivalence*, and *policy-interface equivalence*; their failure to coincide defines *reward-interface dependence*. Our audit combines closed-loop interventions, which measure behavioral consequences, with teacher-forced replay, which isolates the direct effect of reward on the next action. Zero and donor controls then ask whether coordinate sensitivity is tied to total reward use or to trajectory-specific reward evidence.

Across five independently trained AMAGO policies in each of two continuous-control families, matched reward-coordinate interventions change actions and original-task return. HalfCheetah exhibits a non-monotone response that shifts with the training coordinate; Ant-Fwd-Back reproduces a direct and reversible reward-to-action pathway under different morphology and task semantics. Teacher-forced replay places the first action change before environmental feedback. In HalfCheetah, activation patching further shows that the decoded target axis does not dominate the causal effect, whereas structured mismatch subspaces recover much of it as their dimension grows. A recurrent RL^2 carrier extends sequential reward use across architecture and task, while Decision Transformer controls expose a boundary: coordinate sensitivity is weaker when actions use little trajectory-specific reward evidence.

Our contributions are:

1. a formal separation of objective, information, and policy-interface equivalence, together with exact and approximate conditions for interface stability;
2. a matched causal audit that connects reward-coordinate changes to direct action effects, closed-loop consequences, and internal representations across two continuous-control families; and
3. cross-carrier evidence that reward-interface dependence is organized by a policy's reliance on trajectory-specific reward evidence.

2 Reward Equivalence for Learned Adaptation

2.1 Two reward channels

Let $z \sim p(z)$ denote a latent task and let a history-conditioned policy act according to

$$\pi_\theta(a_t \mid h_t), \qquad h_t = (o_{\le t}, a_{<t}, r^{\mathrm{act}}_{<t}, d_{<t}). \tag{1}$$

The environment reward r^{env}_t defines the reported objective, a possibly transformed r^{opt}_t supplies learning targets, and r^{act}_t enters the policy history. Ordinary training often sets all three equal. Our evaluation freezes θ and r^{env}, then intervenes only on r^{act}; there is no optimizer at test time. This separation prevents actor-interface effects from being attributed to changed gradients or objectives.

2.2 Three notions of equivalence

Consider a coordinate map $g : \mathbb{R} \to \mathbb{R}$ applied to each actor-visible reward, and let $T_g(h_t)$ denote the resulting history.

Definition 2.1 (Objective equivalence)**.** Two reward descriptions are objective equivalent on a task family if they induce the same ordering over policies under the stated horizon and termination assumptions.

Definition 2.2 (Information equivalence)**.** Two actor-visible reward histories are information equivalent if they generate the same information about the latent task, e.g., $\sigma(r_{<t}) = \sigma(g(r_{<t}))$ for bijective component-wise g.

Definition 2.3 (Policy-interface equivalence)**.** A fixed policy is interface equivalent under g on a history distribution $\mathcal{H}$ if

$$\pi_\theta(\cdot \mid h) = \pi_\theta(\cdot \mid T_g(h)) \quad \text{for } \mathcal{H}\text{-almost every } h. \tag{2}$$

When objective and information equivalence hold but policy-interface equivalence fails, we call the resulting sensitivity *reward-interface dependence*. This property is defined relative to a policy, coordinate family, and history distribution.

Proposition 2.4 (Objective and information equivalence do not imply interface equivalence)**.** *For a fixed-horizon task, let $g_{a,b}(r) = ar + b$ with $a > 0$. The transformed objective return satisfies $J_g^{\mathrm{obj}}(\pi) = aJ^{\mathrm{obj}}(\pi) + bH$, so policy ordering is preserved, and g is bijective, so the reward histories carry the same information. Neither property implies $\pi_\theta(\cdot \mid h) = \pi_\theta(\cdot \mid T_g(h))$ for a fixed learned policy.*

The first two statements follow from fixed horizon and invertibility. For the last, choose any h with $T_g(h) \neq h$ and a history function that selects different actions on those inputs. Accordingly, objective-level invariance leaves the learned extension beyond the training interface unspecified. Interface equivalence requires an additional property—architectural equivariance, canonicalization, or coverage—that classical reward invariance does not supply. Positive offsets need not preserve ordering under policy-dependent termination, so our offset claims use fixed length or the absorbing-state convention in Section B.

The separation result also suggests constructive conditions under which the missing implication can be recovered. Let $\mathcal{G}$ be a class of coordinate maps and call C a *$\mathcal{G}$-canonicalizer* if $C(T_g(h)) = C(h)$ for every $g \in \mathcal{G}$ and history h in its domain.

Proposition 2.5 (Canonicalization suffices)**.** *If $\pi_\theta(\cdot \mid h) = \bar{\pi}_\theta(\cdot \mid C(h))$ for a $\mathcal{G}$-canonicalizer C, then π_θ is policy-interface equivalent under every $g \in \mathcal{G}$.*

Exact canonicalization is stronger than necessary. A direct action effect also yields a finite-horizon behavioral certificate when it is controlled uniformly. Write TV for total variation distance.

Proposition 2.6 (Approximate interface bound)**.** *Consider a horizon-H environment with $|r_t^{\mathrm{env}}| \leq R_{\max}$. Suppose that, at every time t, any physically matched native and g-transformed prefixes reachable before their first action disagreement satisfy*

$$\mathrm{TV}(\pi_\theta(\cdot \mid h_t), \pi_\theta(\cdot \mid T_g(h_t))) \leq \epsilon_t. \tag{3}$$

Then

$$|J(\pi; \mathrm{id}) - J(\pi; g)| \leq 2R_{\max} \sum_{t=0}^{H-1} (H - t)\epsilon_t. \tag{4}$$

In particular, a uniform ϵ gives an $R_{\max}H(H+1)\epsilon$ bound.

The proof couples actions maximally until their first disagreement and bounds the remaining reward range; Section A gives details. Proposition 2.6 is a distributional certificate based on prefixwise TV, whereas the carrier-specific D_g metrics below diagnose direct interface effects. We report D_g with closed-loop Δ_g to relate local sensitivity to its realized consequence.

2.3 Competence-constrained robustness

A policy that ignores reward is trivially interface invariant but cannot adapt using reward evidence. Let $J(\pi; g)$ be original-task return when only the actor sees $g(r)$, and let d compare action distributions

on matched histories. We report

$$\Delta_g = \mathbb{E}[J(\pi; \mathrm{id}) - J(\pi; g)], \tag{5}$$

$$D_g = \mathbb{E}_{h\sim\mathcal{H}} d(\pi_\theta(\cdot \mid h), \pi_\theta(\cdot \mid T_g(h))), \tag{6}$$

$$\rho_g = D_g / D_0, \qquad D_0 > 0, \tag{7}$$

where D_0 is the corresponding direct effect of masking the coordinate. Here Δ_g measures closed-loop coordinate regret, D_g is the direct fixed-history effect, and ρ_g expresses mismatch as a fraction of total coordinate reliance. We pair these quantities with native competence and absolute $J(\pi; g)$ so that a policy cannot appear robust merely by performing poorly or ignoring reward.

Observable implications. A nonzero D_g on a matched physical history rejects policy-interface equivalence before trajectories diverge; a nonzero Δ_g establishes a consequence under the original objective. The ratio ρ_g places that mismatch relative to the policy's total reward reliance. Training-coordinate interactions then distinguish learned calibration from magnitude sensitivity, while recovery after a finite pulse distinguishes ongoing interface use from a lasting disturbance. These implications determine the audit sequence below.

3 A Matched Reward-Interface Audit

Figure 2 summarizes the audit sequence: establish native competence, apply matched actor-visible reward edits, and separate rollout consequences, direct action effects, and trajectory-specific reward reliance.

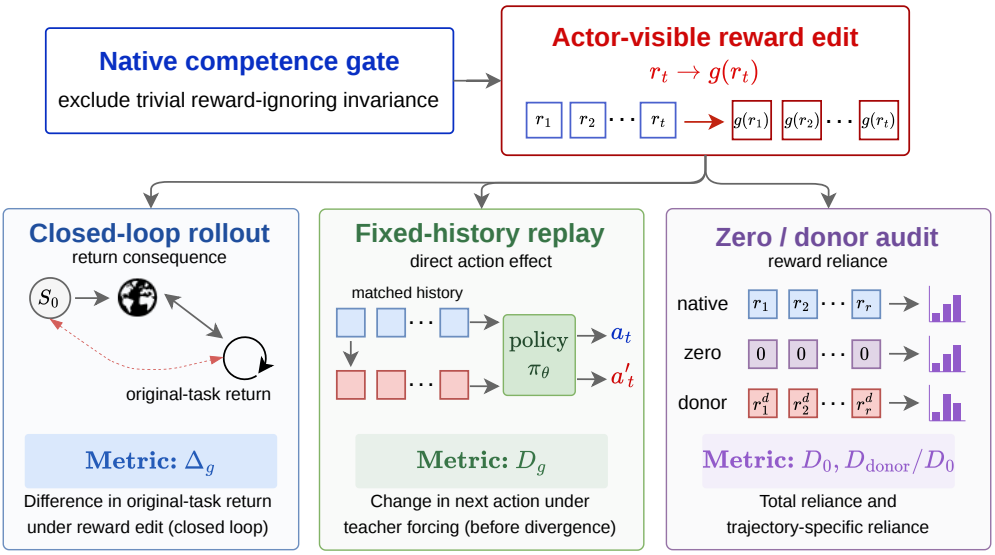


Figure 2: **Matched reward-interface audit.** Paired rollouts measure original-return consequences, fixed-history replay isolates direct action effects, and zero/donor controls separate total from trajectory-specific reward reliance.

3.1 Continuous-control carriers and capability gates

Our primary agent is AMAGO (Grigsby et al., 2024a), a three-layer Transformer trained on HalfCheetah target-velocity meta-RL. A hidden target persists across three inner episodes and must be inferred while memory continues through soft resets. Five independently trained checkpoints satisfy a native adaptation criterion. For a second morphology and task, we train five further checkpoints on an Ant forward/backward family following Rakelly et al. (2019), using Gymnasium's Ant-v4 (Towers et al., 2026). Its hidden goal again persists for three episodes; fixed 200-step episodes and disabled unhealthy termination preserve objective equivalence for offsets. Each Ant policy satisfies native competence and reward-reliance criteria. Training seeds are the inferential units; architecture, thresholds, and panel construction appear in Section B.

3.2 MATCHED INTERVENTIONS

At frozen weights we replace only actor-visible reward with $0.5r$, $0.75r$, $1.25r$, $2r$, $r - 0.5$, or $r + 0.5$. Dynamics, evaluation reward, latent targets, policy parameters, and paired randomness remain unchanged. Static edits test a coordinate throughout adaptation; switch edits introduce it within a trajectory; finite pulses restore native reward after a bounded interval and test reversibility.

Closed-loop rollouts estimate Δ_g, but later differences combine the direct policy response with feedback through altered states. Teacher-forced replay instead holds the native observation–action history fixed, edits only actor-visible reward, and measures D_g before the action can affect the environment. Agreement between replay and rollout connects a direct path to its realized consequence; pulse recovery distinguishes ongoing interface use from a lasting disturbance.

3.3 RELIANCE AND CROSS-CARRIER CONTROLS

The *zero mask* removes the reward-like coordinate and measures total reliance. The *donor mismatch* inserts a real reward path from another trajectory while holding physical history fixed; values remain realistic, but their state–action correspondence is broken. Donor effect relative to zero effect measures trajectory-specific use rather than sensitivity to reward level alone. We repeat these controls in five independently trained RL^2 GRUs on a latent two-arm bandit and in public medium Decision Transformers on HalfCheetah, Hopper, and Walker2d. This changes architecture, action type, and reward distribution while preserving the central question: does the policy use sequential reward evidence, and do equivalent coordinates perturb that use? Full protocols and replay checks appear in Section B.

4 EQUIVALENT COORDINATES ALTER COMPETENT ADAPTATION

4.1 A DIRECTIONAL, NON-MONOTONE OPERATING RANGE

On HalfCheetah, Figure 3a shows a directional, non-monotone response. Downward scaling and a positive offset impair control, a moderate upward scale improves it, and a larger upward scale is again harmful. All five training seeds and stochastic-policy repeats share these directions. The learned interface has an operating range rather than a binary reward detector or a monotone "farther is worse" response. Section C reports the full panel.

The response also follows the coordinate used during training. Under $2r$, ordinary-native policies lose 120.94 ± 29.17 return, whereas doubled-coordinate-native policies lose only 3.74 ± 8.94; the offset comparison yields an ordinary-minus-shaped loss difference of 61.14 ± 20.05. Both paired interactions have the same direction in every training seed. This alignment identifies a learned reward coordinate rather than sensitivity to absolute input magnitude alone.

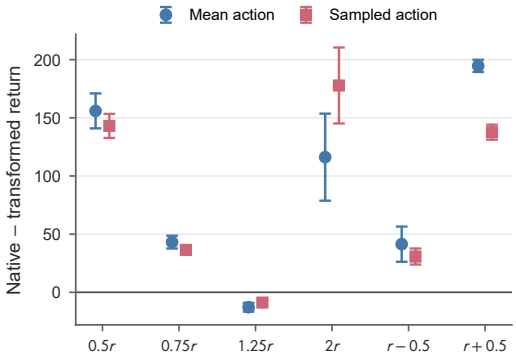
4.2 STATIC MISMATCH AND REVERSIBLE PULSES

On HalfCheetah, activating a coordinate from the beginning reproduces the same dose response, showing that an abrupt switch is unnecessary. During finite pulses, performance follows the active coordinate and returns close to baseline after native reward is restored; the temporary benefit of $1.25r$ recedes in the same way. Coordinate effects track online calibration rather than permanent state corruption (Figure 3b).

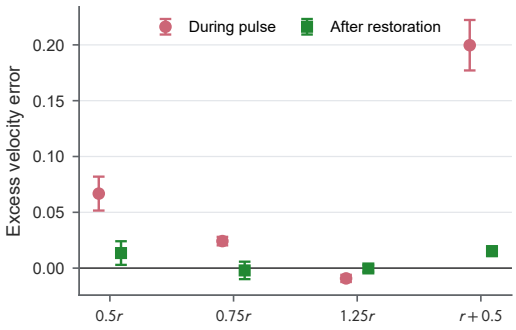
4.3 THE RESPONSE SURVIVES POLICY SAMPLING

On HalfCheetah, deterministic mean actions and sampled actions give the same ordering and all six effect directions. Sampling changes effect size, but not the non-monotone operating point. The ordering persists under deployed action distributions rather than arising from mean-action evaluation.

A second continuous-control task family. On Ant-Fwd-Back, all five policies use reward evidence: zeroing actor-visible reward yields an original-return loss of 494.1 ± 55.8 (mean $\pm$ SE). The affine audit reproduces coordinate sensitivity: $0.5r$ and $2r$ yield static losses of 107.9 ± 7.3 and 224.7 ± 70.1, respectively, in all five seeds, with closely matched switch effects. Under frozen physical histories,



Coordinate regret. Native minus transformed return after a transition-50 switch, under mean and sampled actions.



Pulse and recovery. Excess velocity error during a 50-step edit and after native reward is restored.

Figure 3: **Equivalent reward coordinates induce a non-monotone operating range.** Points are means and error bars are SE across five independently trained policies.

each of the six coordinates changes the first eligible action in all $80/80$ trajectories under both static and switch schedules. Restoring native reward reduces late action displacement by 93.2–94.9% across three pulse coordinates. The direct, reversible pathway extends across morphology and task semantics (Section D).

5 The Reward Token Has a Direct Path to Action

5.1 Direct Fixed-History Causality

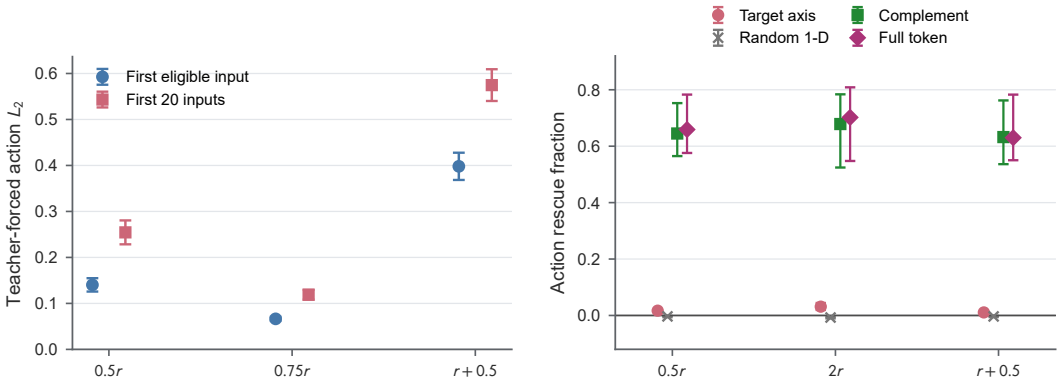
Closed-loop evaluation establishes practical harm but confounds later reward effects with changed physical trajectories. The frozen-history audit removes this ambiguity. Native replay reproduces every stored deterministic action exactly in both continuous-control families. For each family, across all six affine coordinates and five training seeds, the first eligible reward edit changes the next action in every audited trajectory under both static and switch schedules.

The displacement is already present at the first eligible action and grows as edited evidence accumulates. In HalfCheetah, its magnitude is graded across coordinates, and the same ordering appears in an untouched three-seed holdout (Figure 4a). This localizes a direct reward-token-to-action pathway before physical divergence.

A HalfCheetah single-token audit isolates the marginal reward effect by editing only the current token at each native prefix. Four transformations overlap the central 98% of native audit rewards by 94.6–98.1%; within that support, action displacement remains $.061$–$.281$ and is nonzero for every policy. The direct pathway appears without accumulated counterfactual history or a reward-range excursion (Table 5).

5.2 From a Semantic-Axis Failure to a Dimensionality Audit

In HalfCheetah, a cross-fit linear probe decodes target velocity almost perfectly, and affine histories retain decodability while shifting its calibration. Predictive accuracy alone does not establish use (Hewitt & Liang, 2019; Elazar et al., 2021; Belinkov, 2022); we patch activations after two Transformer layers following causal-abstraction and interchange-intervention logic (Geiger et al., 2021; 2022). Correcting the one-dimensional semantic axis rescues only 0.01–0.04 of the action effect, comparable to matched random directions. Its complement rescues 0.52–0.78, close to a full-current-token patch (Figure 4b). That panel averages two layer boundaries and immediate/late windows; ranges span four transform–window cells on three holdout seeds. The decoded target axis explains little of the action effect, motivating a broader subspace analysis.



Direct displacement. Teacher-forced reward edits alter the first eligible action and grow over the first 20 edited inputs.

Causal rescue. Correcting the decoded target axis explains little; complement and full-token patches recover most of the effect.

Figure 4: **Reward edits act through the policy before environment feedback, but not through the decoded target axis alone.** All interventions hold the physical history fixed.

We next cross-fit nested subspaces on held-out simulator-seed groups in three untouched policies. Full-current-token patches rescue 0.626–0.809 of the action effect across transform–window means. Mismatch PCA reaches 80% of this ceiling with $k = 8$–32 in every transform–window–policy cell; at $k = 32$, it rescues 0.609–0.771, versus 0.074–0.109 for matched random spaces. Action-supervised PLS-1 reaches the criterion in $0/18$ cells and PLS-6 in $1/18$. Figure 5 normalizes rescue by the full-token ceiling and averages six transform–window cells within policy; error bars are SE across three holdout policies. Under these cross-fitted constructions, the effect is structured but is not dominated by a stable global 1–6D linear mediator; the audit does not assign an intrinsic representation dimension.

Together, fixed-history replay and patching connect the reward edit, the current representation, and the next action. The action changes before physical divergence, while restoring structured representation components rescues part of the native decision. Information-equivalent histories can therefore induce different computations because the learned encoder and action readout are calibrated to a particular reward coordinate.

6 Coordinate Sensitivity Tracks Reward Reliance

6.1 Meta-RL uses sequential reward evidence

Table 1 reports fixed-history effects. Zeroing feedback strongly changes both AMAGO and RL^2 decisions; a realistic donor reward path is nearly as disruptive when mismatched with the recipient's states and actions. Both agents therefore use trajectory–reward correspondence, not merely global reward level. The competent, independently trained RL^2 carrier extends this result across architecture, action type, task, and reward distribution.

Zero feedback also sharply degrades AMAGO return and RL^2 optimal-action rate. In both carriers, the primary affine edits exceed half of the zero-mask action effect, perturbing a substantial fraction of an actively used channel.

The normalization separates reward reliance from coordinate robustness. D_0 asks whether feedback affects the policy; ρ_g asks how strongly a coordinate change perturbs that active pathway. Donor/zero further distinguishes trajectory-specific evidence from a static target level. This two-axis view supports comparison across carriers even when their action spaces require different direct-effect metrics.

Table 1: Fixed-history coordinate reliance. Metrics are carrier specific; ratios normalize within carrier. Uncertainty is SE across five training seeds for AMAGO and RL^2.

Carrier	Zero	Donor	Donor/zero
AMAGO-HalfCheetah (action L_2)	$.7167 \pm .0443$	$.6297 \pm .0308$	$.884 \pm .036$
RL^2 (policy TV)	$.4854 \pm .0020$	$.3524 \pm .0071$	$.726 \pm .012$
DT-HalfCheetah (RMSE)	.00977	.00047	.048
DT-Hopper (RMSE)	.0947	.00608	.064
DT-Walker2d (RMSE)	.0220	.00064	.029

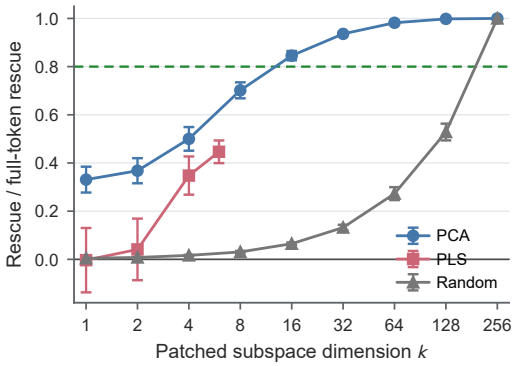


Figure 5: **Structured subspaces rescue action.** Rescue is normalized by the full-token patch; the dashed line marks 80%.

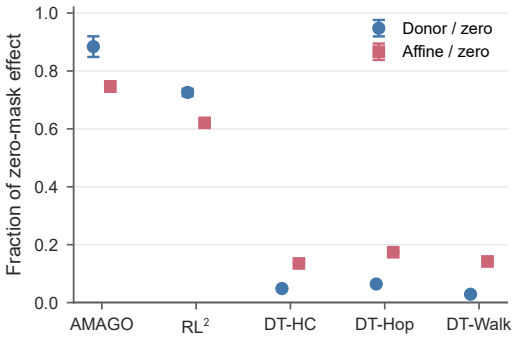


Figure 6: **Sensitivity follows trajectory-specific reward use.** Effects are normalized by the zero-mask intervention.

6.2 Decision Transformer exposes a boundary

The public Decision Transformers behave differently. Donor RTG paths change almost no actions, indicating little use of trajectory-specific updates. Zeroing RTG has little effect in HalfCheetah but causes directionally consistent losses in Hopper and Walker2d: these checkpoints use target level without substantially using the update path probed by donor mismatch.

This contrast defines the mechanism's scope: Figure 6 shows that affine effects occupy only a small fraction of the DT zero-mask effect, unlike in both meta-RL carriers. Coordinate dependence is most pronounced when the policy uses reward as trajectory-specific evidence for adaptation. Direct action displacement and rollout loss remain distinct because dynamics can damp or amplify a local policy change (Section F); the complete audit pairs coordinate reliance with coordinate regret.

7 Related Work

Learned adaptation and reward-conditioned sequence models. Meta-RL treats adaptation as a learned computation (Beck et al., 2025). Recurrent agents encode a fast learning procedure in hidden state (Duan et al., 2016; Wang et al., 2017; Ni et al., 2022); task-inference methods represent uncertainty explicitly (Humplik et al., 2019; Rakelly et al., 2019; Zintgraf et al., 2020); and attention-based agents learn from longer cross-episode histories (Mishra et al., 2018; Melo, 2022; Laskin et al., 2023; Grigsby et al., 2024a;b). Return-conditioned sequence models likewise expose reward-bearing variables to the policy (Chen et al., 2021; Janner et al., 2021; Yamagata et al., 2023; Konan et al., 2023). We ask whether equivalent coordinates induce the same fixed-policy computation.

Reward transformations and normalization. Classical shaping preserves optimal policies under specific transformations (Ng et al., 1999; Wiewiora, 2003), while reward-learning invariances concern objective identifiability (Skalse et al., 2023). PopArt and return-based scaling instead stabilize learning across magnitudes (van Hasselt et al., 2016; Schaul et al., 2021; Sullivan et al., 2023), and

reward transformations calibrate preference models (Wang et al., 2024). We freeze both policy and evaluation objective and change only the actor-visible coordinate.

Generalization, invariance, and underspecification. RL generalization usually varies environments, observations, or tasks (Cobbe et al., 2019; Song et al., 2019; Zhang et al., 2021). We hold the physical task fixed and vary one policy-input coordinate, yielding a structured underspecification test (D'Amour et al., 2022). Because reward carries evidence, invariance must preserve adaptation rather than suppress the channel; Propositions 2.5 and 2.6 give exact and approximate interface targets.

Causal representation analysis. Predictive probes establish decodability, not use (Hewitt & Liang, 2019; Belinkov, 2022). Counterfactual removal and activation interventions connect representations to behavior (Elazar et al., 2021; Geiger et al., 2021; 2022); our cross-fit, dimension-matched patches ask how restoring selected representation components restores action.

8 Discussion and Limitations

The reward coordinate is part of the learned policy. For a history-conditioned agent, reward specifies both an objective and an input used for adaptation. Training-coordinate alignment and reversible pulses show that the second role is learned: changing only actor-visible reward can alter behavior with fixed weights, unchanged dynamics, and the same evaluation objective. A deployed policy is specified not only by its parameters, but also by the coordinate in which its reward history is presented.

A common formal and empirical account. Objective and information equivalence constrain the task and its evidence, but not the fixed policy's computation. Canonicalization supplies an exact route to interface equivalence, and the approximate bound converts uniformly small prefixwise action effects into a return guarantee. The audit reveals the complementary failure mode: matched-history replay identifies a direct reward-to-action effect, rollout interventions establish its consequence, and training-coordinate interactions and pulse recovery show ongoing learned calibration. Representation and carrier comparisons refine the mechanism: the effect is structured beyond the decoded task axis and is most pronounced when reward provides trajectory-specific evidence.

Implications for evaluation and design. Reward-conditioned policies should be documented together with the coordinate presented to the actor. Three complementary tests suffice for a basic audit: invertible coordinate changes probe behavioral equivalence, matched-history replay separates direct policy effects from environmental feedback, and zero/donor interventions measure total and trajectory-specific reward reliance. This combination distinguishes stable adaptation from reward-ignoring invariance and separates local mismatch from realized return cost. Explicit calibration, canonical reward histories, and architectural equivariance are concrete design targets; the augmentation study shows why each must be assessed jointly with native competence (Section H).

Limitations. The two continuous-control families share the AMAGO architecture; RL^2 changes architecture and task type, while Decision Transformer provides a reliance boundary. Representation interventions cover HalfCheetah and cross-fitted global linear subspaces. Five seeds per AMAGO family and three untouched HalfCheetah holdouts establish repeatability rather than precise population uncertainty. A Meta-World pilot did not meet its native adaptation criterion (Section I); broader continuous-control architectures and nonlinear or sample-dependent analyses remain open.

9 Conclusion

Reward transformations that preserve an objective and its information need not preserve the behavior of a learned history-conditioned policy. Across two continuous-control families and a recurrent meta-RL carrier, matched interventions reveal a direct, reversible reward-to-action pathway whose sensitivity tracks the use of trajectory-specific reward evidence. The distinction among objective, information, and policy-interface equivalence makes reward coordinates an explicit property of learned adaptation. Reporting that coordinate and auditing equivalent reparameterizations are practical steps toward policies whose adaptation is stable without discarding informative feedback.

AI USE STATEMENT

In this work, generative AI tools were used to assist with language editing and polishing of the manuscript, literature search, theory check, debugging of research and drawing code, and analysis and interpretation of experimental results. All AI-assisted suggestions, analyses, and outputs were critically reviewed and verified by the authors. The authors determined the final research decisions, methodological choices, experimental conclusions, and manuscript content. The authors take full responsibility for the accuracy, validity and integrity of the work.

ETHICS STATEMENT

This work studies the reliability and interpretability of learned reinforcement-learning interfaces. It involves no human participants or sensitive personal data and introduces no deployed decision-making system.

REPRODUCIBILITY STATEMENT

The main text specifies interventions, inferential units, capability gates, and causal controls. Section A gives complete proofs; Section B reports architectures, budgets, panels, timing, integrity tolerances, public-checkpoint revisions, donor construction, and cross-fitted subspace estimation.

REFERENCES

Jacob Beck, Risto Vuorio, Evan Zheran Liu, Zheng Xiong, Luisa Zintgraf, Chelsea Finn, and Shimon Whiteson. A tutorial on meta-reinforcement learning. *Foundations and Trends® in Machine Learning*, 18(2-3):224–384, April 2025. ISSN 1935-8245. doi: 10.1561/2200000080. URL http://dx.doi.org/10.1561/2200000080.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, March 2022. doi: 10.1162/coli_a_00422. URL https://aclanthology.org/2022.cl-1.7/.

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Michael Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling, 2021. URL https://arxiv.org/abs/2106.01345.

Karl Cobbe, Oleg Klimov, Chris Hesse, Taehoon Kim, and John Schulman. Quantifying generalization in reinforcement learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1282–1289. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/cobbe19a.html.

Alexander D'Amour, Katherine Heller, et al. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022. URL http://jmlr.org/papers/v23/20-1335.html.

Yan Duan, John Schulman, Xi Chen, Peter L. Bartlett, Ilya Sutskever, and Pieter Abbeel. Rl^2: Fast reinforcement learning via slow reinforcement learning, 2016. URL https://arxiv.org/abs/1611.02779.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021. doi: 10.1162/tacl_a_00359. URL https://aclanthology.org/2021.tacl-1.10/.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 9574–9586. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/4f5c422f4d49a5a807eda27434231040-Paper.pdf.

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah Goodman, and Christopher Potts. Inducing causal structure for interpretable neural networks. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 7324–7338. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/geiger22a.html.

Jake Grigsby, Jim Fan, and Yuke Zhu. Amago: Scalable in-context reinforcement learning for adaptive agents. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 26919–26952, 2024a. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/7204434dcb9383a1454dc1e97e58ea9c-Paper-Conference.pdf.

Jake Grigsby, Justin Sasek, Samyak Parajuli, Daniel Adebi, Amy Zhang, and Yuke Zhu. Amago-2: Breaking the multi-task barrier in meta-reinforcement learning with transformers, 2024b. URL https://arxiv.org/abs/2411.11188.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

Jan Humplik, Alexandre Galashov, Leonard Hasenclever, Pedro A. Ortega, Yee Whye Teh, and Nicolas Heess. Meta reinforcement learning as task inference, 2019. URL https://arxiv.org/abs/1905.06424.

Michael Janner, Qiyang Li, and Sergey Levine. Offline reinforcement learning as one big sequence modeling problem, 2021. URL https://arxiv.org/abs/2106.02039.

Sachin G. Konan, Esmaeil Seraj, and Matthew Gombolay. Contrastive decision transformers. In Karen Liu, Dana Kulic, and Jeff Ichnowski (eds.), *Proceedings of The 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pp. 2159–2169. PMLR, 14–18 Dec 2023. URL https://proceedings.mlr.press/v205/konan23a.html.

Michael Laskin, Luyu Wang, Junhyuk Oh, Emilio Parisotto, Stephen Spencer, Richie Steigerwald, DJ Strouse, Steven Stenberg Hansen, Angelos Filos, Ethan Brooks, maxime gazeau, Himanshu Sahni, Satinder Singh, and Volodymyr Mnih. In-context reinforcement learning with algorithm distillation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=hy0a5MMPUv.

Luckeciano C Melo. Transformers are meta-reinforcement learners. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 15340–15359. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/melo22a.html.

Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. A simple neural attentive meta-learner. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=B1DmUzWAW.

Andrew Y. Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In Ivan Bratko and Saso Dzeroski (eds.), *Proceedings of the Sixteenth International Conference on Machine Learning (ICML 1999), Bled, Slovenia, June 27 - 30, 1999*, pp. 278–287. Morgan Kaufmann, 1999.

Tianwei Ni, Benjamin Eysenbach, and Ruslan Salakhutdinov. Recurrent model-free RL can be a strong baseline for many POMDPs. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16691–16723. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/ni22a.html.

Kate Rakelly, Aurick Zhou, Chelsea Finn, Sergey Levine, and Deirdre Quillen. Efficient off-policy meta-reinforcement learning via probabilistic context variables. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5331–5340. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/rakelly19a.html.

Tom Schaul, Georg Ostrovski, Iurii Kemaev, and Diana Borsa. Return-based scaling: Yet another normalisation trick for deep rl, 2021. URL https://arxiv.org/abs/2105.05347.

Joar Max Viktor Skalse, Matthew Farrugia-Roberts, Stuart Russell, Alessandro Abate, and Adam Gleave. Invariance in policy optimisation and partial identifiability in reward learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 32033–32058. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/skalse23a.html.

Xingyou Song, Yiding Jiang, Stephen Tu, Yilun Du, and Behnam Neyshabur. Observational overfitting in reinforcement learning, 2019. URL https://arxiv.org/abs/1912.02975.

Ryan Sullivan, Akarsh Kumar, Shengyi Huang, John Dickerson, and Joseph Suarez. Reward scale robustness for proximal policy optimization via dreamerv3 tricks. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 1352–1362. Curran Associates, Inc., 2023. doi: 10.52202/075280-0067. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/04f61ec02d1b3a025a59d978269ce437-Paper-Conference.pdf.

Mark Towers, Ariel Kwiatkowski, John U. Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Kallinteris Andreas, Markus Krimmel, Arjun KG, Rodrigo De Lazcano Perez-Vicente, J K Terry, Andrea Pierré, Sander V Schulhoff, Jun Jet Tai, Hannah Tan, and Omar G. Younis. Gymnasium: A standard interface for reinforcement learning environments. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2026. URL https://openreview.net/forum?id=qPMLvJxtPK.

Hado van Hasselt, Arthur Guez, Matteo Hessel, Volodymyr Mnih, and David Silver. Learning values across many orders of magnitude, 2016. URL https://arxiv.org/abs/1602.07714.

Jane X Wang, Zeb Kurth-Nelson, Dhruva Tirumala, Hubert Soyer, Joel Z Leibo, Remi Munos, Charles Blundell, Dharshan Kumaran, and Matt Botvinick. Learning to reinforcement learn, 2017. URL https://arxiv.org/abs/1611.05763.

Zihao Wang, Chirag Nagpal, Jonathan Berant, Jacob Eisenstein, Alexander Nicholas D'Amour, Sanmi Koyejo, and Victor Veitch. Transforming and combining rewards for aligning large language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 51161–51176. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/wang24ay.html.

E. Wiewiora. Potential-based shaping and q-value initialization are equivalent. *Journal of Artificial Intelligence Research*, 19:205–208, September 2003. ISSN 1076-9757. doi: 10.1613/jair.1190. URL http://dx.doi.org/10.1613/jair.1190.

Taku Yamagata, Ahmed Khalil, and Raul Santos-Rodriguez. Q-learning decision transformer: Leveraging dynamic programming for conditional sequence modelling in offline RL. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 38989–39007. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/yamagata23a.html.

Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In Leslie Pack Kaelbling, Danica Kragic, and Komei Sugiura (eds.), *Proceedings of the Conference*

on Robot Learning, volume 100 of *Proceedings of Machine Learning Research*, pp. 1094–1100. PMLR, 30 Oct–01 Nov 2020. URL https://proceedings.mlr.press/v100/yu20a.html.

Amy Zhang, Rowan McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction, 2021. URL https://arxiv.org/abs/2006.10742.

Luisa Zintgraf, Kyriacos Shiarlis, Maximilian Igl, Sebastian Schulze, Yarin Gal, Katja Hofmann, and Shimon Whiteson. Varibad: A very good method for bayes-adaptive deep rl via meta-learning, 2020. URL https://arxiv.org/abs/1910.08348.

A Proofs for the Interface Conditions

Proof of Proposition 2.5. For any $g \in \mathcal{G}$, $\pi_\theta(\cdot \mid T_g(h)) = \bar{\pi}_\theta(\cdot \mid C(T_g(h))) = \bar{\pi}_\theta(\cdot \mid C(h)) = \pi_\theta(\cdot \mid h)$. □

Proof of Proposition 2.6. Couple the native and transformed policies maximally at each step while their physical histories remain matched, and use common environment randomness whenever their actions agree. Conditional on no earlier action disagreement, the probability of a first disagreement at time t is at most ϵ_t. If it occurs, the largest possible difference in the remaining evaluation return is $2R_{\max}(H-t)$. Summing over possible first-disagreement times gives

$$|J(\pi; \mathrm{id}) - J(\pi; g)| \le 2R_{\max} \sum_{t=0}^{H-1} (H-t)\epsilon_t. \tag{8}$$

Substituting $\epsilon_t \le \epsilon$ and summing $H + (H-1) + \cdots + 1$ gives $R_{\max} H(H+1)\epsilon$. □

B Experimental Details

AMAGO training and environment. We use AMAGO 3.4.0 at pinned commit `54c25ab6d`. The timestep encoder consumes the observation (including a soft-reset flag), previous action, and previous actor-visible reward; the trajectory encoder also receives the time index. No separate goal field is provided, and the target velocity is not included in the policy input. The policy is a three-layer, width-256 Transformer with rotary position embeddings. Each condition is trained for 100 epochs with 12 asynchronous actors, approximately 1.2 million actor steps, and 100,000 learner updates. Ten fixed target velocities crossed with five simulator seeds form each formal 50-cell panel. The target persists through three 200-step inner episodes while Transformer memory continues through soft resets.

Ant replication. The Ant policies use the same three-layer, width-256 Transformer and 100-epoch, 12-actor training budget. A hidden forward/backward goal persists through three 200-step inner episodes and is absent from the policy input; reward is directional velocity plus the native healthy reward minus control cost. Each inner episode has a fixed 200-step horizon, with unhealthy termination disabled; an additive reward offset therefore contributes the same policy-independent constant under the corresponding objective transformation. Each paired panel crosses two directions with eight simulator seeds. We train seeds $11, 23, 37, 53, 71$. Before affine auditing, each policy must have a positive lower bootstrap bound on native post-probe directional velocity in both directions and on its paired native-minus-zero difference.

Inferential units. For AMAGO and RL^2, independently trained checkpoints are the inferential units. Matched target/simulator cells, trajectories, held-out bandit trials, and policy-sampling repetitions are averaged within checkpoint before computing standard errors. We emphasize effect sizes, standard errors, and seed-direction counts because five policies do not support precise tail inference. The five primary HalfCheetah AMAGO coordinates were fixed before the sampled-action replication; $r - 0.5$ is descriptive. Decision Transformer uses one pinned public checkpoint per environment; its 20 reset seeds quantify paired rollout variability rather than training variability.

Integrity checks. Native teacher-forced replay must reproduce stored actions with maximum error at most 10^{-4} for AMAGO, probabilities at most 10^{-7} for RL^2, and actions at most 10^{-6} for Decision Transformer. All replay checks pass with zero observed error.

For the rescue-versus-dimension audit, each simulator seed is held out in turn when fitting the mismatch PCA or action-supervised PLS basis. Four independently sampled nested random bases are matched by dimension. Basis orthonormal error must be at most 10^{-4}, native feature replay at most 10^{-6}, and the 256-dimensional endpoint must reproduce the exact full-current action within 10^{-4}. Across the three policy seeds, the observed maxima are 9.54×10^{-7}, 0, and 0, respectively; all outputs are finite.

Intervention timing. Environment reward from transition s enters the actor on input $s + 1$. A transition-50 edit therefore first affects action 51. Switch panels retain the transformed coordinate for the remainder of the 600-step meta-trial; pulse panels restore native reward at transition 100.

Cross-carrier protocols. The RL^2 carrier is a 48-unit GRU trained for 300 PPO updates on 12-step Bernoulli two-arm trials. Training gaps are $\pm\{0.15, 0.20, 0.25, 0.30\}$ and held-out gaps are $\pm\{0.175, 0.225, 0.275\}$. Formal seeds 1–5 each use 65,536 common-random-number held-out trials. The Decision Transformer carriers are the public `medium` HalfCheetah-v4, Hopper-v4, and Walker2d-v4 checkpoints at fixed revisions, with targets 6000, 3600, and 5000, context length 20, and return scale 1000. For $g(r) = ar + b$, their displayed RTG is $G'_t = aG_t + b(T - t)$ with $T = 1000$. Early termination uses absorbing-state padding, making bT policy independent. Reset seeds 0–19 are paired across conditions.

Donor construction. For AMAGO, a recipient target receives the reward stream from the target five positions away under the same simulator seed. In RL^2, donor reward comes from a held-out task with the opposite latent-task sign. For Decision Transformer, reset seed i receives the native RTG path from seed $(i + 1) \bmod 20$. These deterministic mappings were frozen before the donor outputs were inspected.

C FULL AMAGO COORDINATE RESULTS

Table 2: Ordinary-native transition-50 coordinate loss under sampled and deterministic mean actions. Loss is native minus transformed original return; uncertainty is SE across five independently trained checkpoints.

Coordinate	Sampled	Mean action	Seed direction
$0.5r$	143.04 ± 10.34	155.96 ± 15.04	5/5 harmful
$0.75r$	36.48 ± 3.49	43.15 ± 5.51	5/5 harmful
$1.25r$	-8.83 ± 2.58	-12.83 ± 3.78	5/5 beneficial
$2r$	177.80 ± 32.67	116.17 ± 37.42	5/5 harmful
$r - 0.5$	30.77 ± 6.90	41.39 ± 15.13	5/5 harmful
$r + 0.5$	137.64 ± 6.48	194.66 ± 5.26	5/5 harmful

Training-coordinate comparisons use separately trained native-coordinate policies. For $2r$, ordinary-native loss is 120.94 ± 29.17, versus 3.74 ± 8.94 for doubled-coordinate-native policies. For $r - 0.5$, the ordinary-minus-shaped loss difference is 61.14 ± 20.05. Both paired contrasts have the same direction in every training seed.

Table 3: Evidence map for the mechanism claim. Each stage changes one source of variation while preserving the quantities identified at earlier stages.

Claim	Identifying intervention	Evidence
The actor learns a reward coordinate	Two five-seed AMAGO task families; static, switch, and pulse panels; HalfCheetah training-coordinate and sampling controls	Sensitivity repeats across morphologies; HalfCheetah effects follow the active and training coordinates.
Reward directly affects action	Teacher-forced replay of the same native physical history in both task families	The first eligible action changes before policy feedback can alter the environment.
The direct effect is structured beyond the decoded target axis	Cross-fit semantic patches and nested PCA, PLS, random, and full-token interventions	The semantic axis explains little; the tested global 1–6D constructions lack consistent rescue, while mismatch PCA reaches 80% rescue with 8–32 dimensions.
Coordinate mismatch is reliance conditional	Zero and donor audits in AMAGO, RL^2, and Decision Transformer	Sequential evidence use tracks coordinate sensitivity; target-level use and under-use produce different signatures.

D ANT FORWARD–BACKWARD REPLICATION

All five checkpoints pass the competence and zero-mask reliance gates. Removing actor-visible reward produces an original-return loss of 494.10 ± 55.76 and a post-probe directional-velocity loss of $.961 \pm .110$, both positive in every training seed. Table 4 reports the affine panel.

Table 4: Ant-Fwd-Back reward-coordinate effects. Return loss is native minus transformed original return. D_g is the teacher-forced first-eligible-action L_2 distance under a static transformation. Values are mean $\pm$ SE across five independently trained policies.

Coordinate	Static loss	Switch-50 loss	First-action D_g
$0.5r$	107.9 ± 7.3	93.0 ± 15.2	$.268 \pm .034$
$0.75r$	29.1 ± 10.6	28.3 ± 10.5	$.201 \pm .028$
$1.25r$	12.4 ± 30.3	20.7 ± 22.2	$.134 \pm .018$
$2r$	224.7 ± 70.1	210.3 ± 66.0	$.278 \pm .034$
$r - 0.5$	41.2 ± 15.7	58.0 ± 15.8	$.350 \pm .071$
$r + 0.5$	77.4 ± 50.1	79.5 ± 38.2	$.429 \pm .116$

The static and switch return-loss profiles correlate at .939 across the six coordinates. Both $0.5r$ and $2r$ are harmful in all five seeds under both schedules, whereas the near-native $1.25r$ consequence is small and seed variable. Native teacher-forced replay has zero observed error, and each of the six coordinates changes the first eligible action in all $80/80$ trajectories under both static and switch schedules. For pulses using $0.5r$, $0.75r$, and $r + 0.5$, restoring native reward reduces late action displacement by 94.1%, 94.9%, and 93.2%, respectively. These results reproduce coordinate sensitivity and its direct, reversible path without extending the HalfCheetah representation claims to Ant.

E ADDITIONAL MECHANISM RESULTS

Single-token support audit. The main teacher-forced audit edits the reward coordinate along a replayed history. To isolate a local effect, we additionally clone the native pre-step cache at each eligible input, transform only the current reward token, query the next mean action, and discard the counterfactual cache. For each policy, we define the native audit interval as the first-to-99th percentile of rewards in held-out native trajectories and report action displacement before and after restricting transformed values to that interval.

Table 5: **A single in-range reward token changes the next action.** Coverage is the fraction of transformed values inside each policy's central 98% native audit interval. Values are mean $\pm$ SE across five independently trained policies.

Coordinate	Coverage	All-token D_g	In-range D_g
$0.5r$	$.786 \pm .042$	$.142 \pm .004$	$.167 \pm .007$
$0.75r$	$.946 \pm .009$	$.069 \pm .001$	$.070 \pm .002$
$1.25r$	$.981 \pm .001$	$.063 \pm .001$	$.061 \pm .001$
$2r$	$.951 \pm .003$	$.225 \pm .003$	$.196 \pm .003$
$r - 0.5$	$.977 \pm .001$	$.281 \pm .007$	$.281 \pm .007$
$r + 0.5$	$.121 \pm .017$	$.381 \pm .010$	$.446 \pm .023$

The high-overlap coordinates retain direct effects of comparable scale after restriction. The interval comes from held-out native audit trajectories, so this is a marginal audit-range diagnostic rather than a claim about full training-distribution support.

The frozen-history panel contains static and switch variants for all six affine coordinates, plus pulse variants for $0.5r$, $0.75r$, and $r + 0.5$. For all six static/switch coordinates, the first eligible action changes in every audited trajectory across all five training seeds. After reward restoration, discrepancies persist while edited tokens remain in the finite rolling context and subsequently decay to zero when they leave it. This schedule-dependent hysteresis is consistent with reward evidence propagating through the Transformer context rather than permanent corruption.

The semantic causal-closure protocol cross-fits target-velocity probes and evaluates two layer boundaries, immediate and late windows, eight matched-norm random directions, shuffled wrong-target controls, the semantic orthogonal complement, and a full-current-token patch. All five integrity checks pass; the maximum projection-coordinate error is 5.36×10^{-7} and the maximum matched-random norm error is 2.38×10^{-7}.

The native target-velocity probe has RMSE 0.043 and correlation 0.9985. Affine histories preserve high correlation but shift calibration: mean signed biases are -0.136 for $0.5r$, $+0.181$ for $2r$, and -0.289 for $r + 0.5$.

The one-dimensional mechanism gate required semantic rescue of at least 0.10, improvements of at least 0.05 over both matched-random and wrong-target controls, and induction cosine of at least 0.50 in two of three held-out training seeds. All three primary transforms pass in $0/3$ holdout seeds. No primary transform meets the criterion in any of the three holdout seeds, so the tested one-dimensional semantic mediator is not supported.

E.1 Cross-fitted rescue versus dimension

The subsequent dimensionality audit uses the same three untouched policies and physical trajectories but fits all subspaces without the held-out simulator-seed group. The full-current rescue and smallest PCA dimension satisfying both 80% of full rescue and a 0.10 advantage over matched random are reported in Table 6. PLS-1 satisfies the cellwise 80% criterion in $0/18$ transform–window–policy cells; PLS-6 does so in $1/18$. Thus neither tested PLS construction consistently meets the rescue criterion across windows and seeds. PCA succeeds with 8–32 dimensions in every cell, but this should not be read as an intrinsic or nonlinear dimension estimate.

Table 6: Cross-fitted causal-subspace audit at the primary residual boundary. Full rescue is mean$\pm$SE across three holdout policies; PCA-32 and random-32 are mean rescue. k_{80} lists the smallest qualifying PCA dimension for policy seeds 37/53/71.

Coordinate	Window	Full	PCA-32	Random-32	k_{80} by seed
$0.5r$	Immediate	$.783 \pm .031$	.720	.092	16/32/32
$0.5r$	Late	$.670 \pm .098$	.642	.099	8/8/16
$2r$	Immediate	$.768 \pm .011$	.669	.074	32/32/32
$2r$	Late	$.809 \pm .031$	.771	.105	16/16/16
$r + 0.5$	Immediate	$.783 \pm .027$	.732	.086	16/16/16
$r + 0.5$	Late	$.626 \pm .111$	.609	.109	8/8/16

F Descriptive DT Action–Score Association

Across the nine Decision Transformer environment-by-affine cells, fixed-history action RMSE correlates with absolute paired normalized-score change at $r = 0.855$. This association is descriptive: the cells share only three checkpoints and environments, and dynamics can amplify or damp similar local action changes. We include it as a consistency diagnostic rather than evidence for a general predictive law.

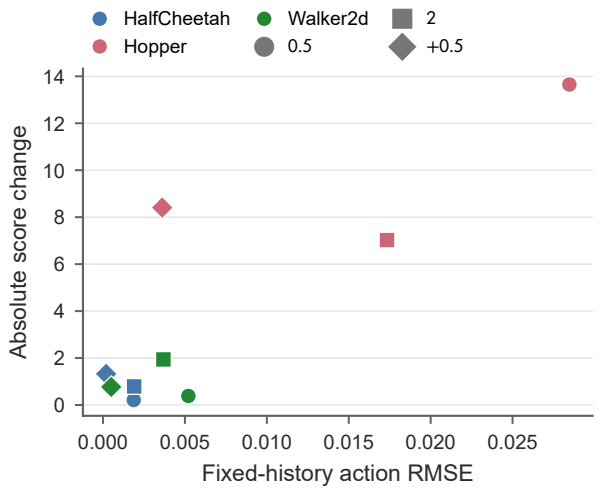


Figure 7: **Decision Transformer action and rollout sensitivity.** Nine environment-by-affine cells show a descriptive association ($r = 0.855$); color denotes environment and marker denotes coordinate.

G Objective, Information, and Stability Separations

Additional POMDP experiments in public T-Maze and RockSample implementations illustrate why objective equivalence, information equivalence, and learning stability are distinct. A privileged potential can preserve telescoping return while leaking hidden task information through actor-visible reward. A recursively history-measurable belief potential preserves the actor's information but can still change the finite-sample learning problem and learned interface. We report these additional separations alongside the primary positive-affine causal audit, whose transformations preserve information by construction.

H Competence-Constrained Augmentation

Uniformly sampling one of seven reward coordinates per meta-trial lowers native competence substantially: at epoch 99, native return falls from approximately -233 for matched ordinary baselines to -588.67, while native velocity error rises from 0.209 to 0.904. A native-heavy mixture passes an exploratory two-seed promotion gate but fails an independent three-seed confirmation: large-offset sensitivity reductions replicate, but holdout native return is lower by 70.08, exceeding the frozen 50-point tolerance. We therefore report augmentation as a diagnostic competence–robustness frontier, not as a proposed method.

I Capability-Gated Meta-World Pilot

We separately explored AMAGO on the Meta-World benchmark (Yu et al., 2020), using the ML1 `reach-v2` family as a candidate second modern continuous-control carrier. The protocol froze a 50-task test panel and required native third-attempt success of at least 0.60, third-minus-first success of at least 0.10, and positive original-return improvement before any reward-coordinate intervention. A selected long-context seed was extended from epoch 149 to epoch 249. At epoch 249 its three attempt-wise success rates were $0.46/0.52/0.54$, third-minus-first success was 0.08, and original return improved by 27.10. Because two native capability criteria failed, the affine panel and independent confirmation seeds were not run. We therefore exclude Meta-World from the coordinate claim rather than interpreting an under-adapting carrier as robust or insensitive.