BEHAVIORALLY SILENT NOW, CAUSALLY ACTIVE LATER: DEFERRED STATE EFFECTS IN RECURRENT POLICIES FOR REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Observation perturbations in reinforcement learning (RL) agents with recurrent policies can have lasting effects on both the policy's internal state and its subsequent behavior. However, when such delayed effects occur, it remains unclear whether they arise from the immediate action change induced by the perturbation or from perturbation information retained in the policy's internal state. To address this, we introduce current action equivalence, which separates these pathways by comparing internal states that produce the same current action. For the one-shot intervention on feedforward policies, reproducing the current action reproduces the downstream effect. In recurrent policies, however, real observation perturbations can induce internal state changes that remain behaviorally silent at the current step yet become increasingly expressed through future actions. We identify these changes as deferred state effects and show through interventions that they causally influence future actions, closed-loop trajectories, and task reward. The effect is consistent across perturbations, recurrent architectures, environments, and independently trained policies. These results show that the effects of a perturbation can remain behaviorally silent at the present step while continuing to shape future decisions in recurrent policies.

1 INTRODUCTION

Reinforcement learning (RL) agents are often evaluated through their observable behavior. When a perturbation no longer affects the current action, its influence may therefore appear to have disappeared (Huang et al., 2017; Pattanaik et al., 2017; Zhang et al., 2020). This intuition is natural for feedforward policies, but it can fail for recurrent agents. A recurrent policy carries internal state across decisions, allowing an observation perturbation to alter what the policy remembers even when that change is weakly expressed through the current action. The agent can therefore appear behaviorally unchanged at the present step while carrying a state change caused by the perturbation that alters later decisions.

This possibility makes delayed perturbation effects causally ambiguous. A later behavioral difference may arise because the perturbation changed the current action and thereby altered the subsequent trajectory. Alternatively, the perturbation may have changed the recurrent state in a way that continues to affect later decisions. Observing a later divergence alone cannot distinguish these two pathways. The key challenge is therefore to separate the effect already mediated by the current action from the effect that remains in the policy state.

We introduce current action equivalence to make this separation possible. Two internal states are current action equivalent when they produce exactly the same current action. By additionally matching the first environment transition, subsequent differences cannot be attributed to different initiating actions or immediate environmental consequences. For the tested one-shot feed-forward intervention, reproducing the current action reproduces the downstream trajectory. Recurrent policies behave differently: even with the same current action and the same first transition, different internal states can still lead to different future actions and trajectories. The important question is whether real observation perturbations actually create internal policy state differences that matter for future

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.

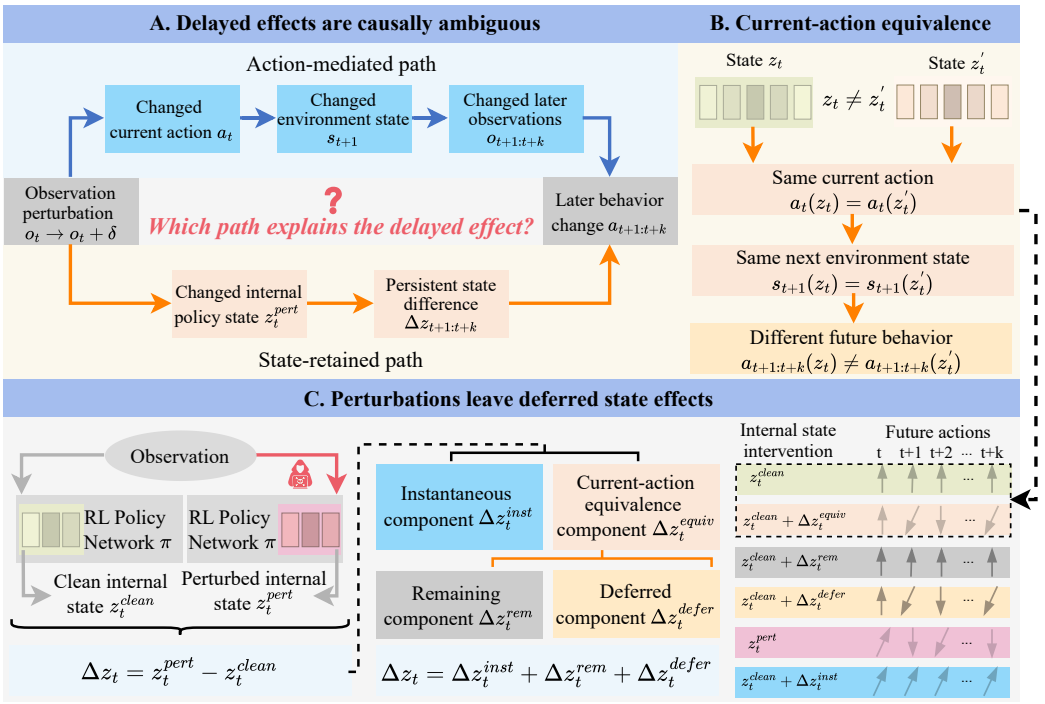

Figure 1: Causal view of delayed perturbation effects. (A) A delayed behavioral change can arise through the current action or through information retained inside the policy. (B) Current action equivalence removes the first explanation by comparing internal states that produce the same current action. (C) Real observation perturbations create a deferred component in internal policy state that is weakly expressed through the current action but becomes more visible later. Controlled interventions show that this component causally affects future actions and closed-loop trajectories.

behavior beyond their effect on the current action. Figure 1 summarizes this causal distinction and the main finding of our study.

The analysis shows that most perturbation-induced state displacement is not yet expressed through the current action. Within this action-equivalent difference, real perturbations exhibit structured components with distinct temporal expression profiles. In particular, one component is weakly expressed at the current step but becomes increasingly visible through future actions; we call this the deferred component. The remaining component shows substantially weaker future action expression.

Causal relevance is then tested by comparing equal norm interventions along the deferred and remaining components from the same perturbed state while preserving the current action. Intervening along the deferred component produces substantially greater recovery of future actions and closed-loop trajectories. Importantly, the same recovery is also reflected in task reward. Moreover, equal norm random directions within the same deferred subspace recover substantially less, while reversing the perturbation-aligned direction moves behavior farther from the clean continuation. These effects remain consistent across observation perturbations, recurrent architectures, control environments, and independently trained policies.

Our contributions are:

- We introduce current action equivalence as a causal criterion for separating effects already expressed through the current action from those that remain active in the internal policy state.
- Our analysis shows that most of the state change caused by a real observation perturbation is not yet expressed through the current action, and that this remaining difference contains structured directions that become more strongly expressed through future actions.

- Intervention experiments, closed-loop evaluations, and task reward analyses validate these deferred state effects, with the same causal pattern consistently observed across diverse recurrent policies and environments.

2 Related Work

Mechanistic analysis of policies. Hidden-activation analyses connect policy representations to learning, planning, and navigation (McInroe et al., 2021; Trim & Grayston, 2024). Interventions manipulate internal features or policy representations (Mini et al., 2023; Taufeeque et al., 2026; Häon et al., 2025), while causal explanations also operate at behavioral or abstract-state levels (Madumal et al., 2020; Lu et al., 2024; Kekić et al., 2026). We instead separate effects mediated by the current action from those that remain active inside the policy.

Activation patching and distributed interventions. Causal tracing, interchange interventions, and distributed alignment test counterfactuals by modifying internal activations (Meng et al., 2022; Geiger et al., 2021; 2024). Their interpretation can depend on corruption, donors, metrics, and interacting mediators (Zhang & Nanda, 2024; Heimersheim & Nanda, 2024; Vaidyanathan et al., 2026; Makelov et al., 2024). Our setting additionally enforces exact current-action equality to isolate effects beyond the present action.

Behavioral subspaces and local output geometry. Low-rank interventions, Jacobian subspaces, and activation subspaces provide tools for identifying behaviorally relevant directions (Geiger et al., 2024; Jiang et al., 2025; Zhu et al., 2021; Sundararajan et al., 2017; Tang et al., 2026). We use such local geometry only within real clean–attacked deviations and verify current-action equality with finite checks.

Recurrent mechanisms and observability. Observability Gramians characterize how state perturbations become visible over time (Kawano & Scherpen, 2021; Kazma & Taha, 2026), while recurrent interventions reveal effects that unfold across steps (Taufeeque et al., 2026). We condition this analysis on current action equivalence and test whether temporally distinct components of real perturbation-induced deviations have different causal effects on future behavior.

3 Comparing Policy States at a Fixed Current Action

3.1 Problem Setting and Current Action Equivalence

We study a frozen deterministic policy at an environment state s_t. From the same state, we get a clean observation o_t and a perturbed observation $\tilde{o}_t$. For a recurrent policy, let z_t denote its internal policy state after processing the observation, and let $a_t(z_t)$ denote the resulting raw action. The clean and perturbed executions produce states z_t^{clean} and z_t^{pert}, respectively. Their difference is the perturbation-induced state deviation that we analyze in the following sections.

Our goal is to distinguish effects already expressed through the current action from effects that remain active inside the policy. We therefore define two internal states z_t and z'_t to be *current-action equivalent* if

$$a_t(z_t) = a_t(z'_t). \tag{1}$$

Current action equivalence is defined on the raw deterministic action. We further require the two executions to have identical clipped actions and to induce the same first environment transition. Thus, if the two executions later produce different actions, the difference cannot be attributed to different initiating actions or first environment transitions.

By fixing the current action in this way, current action equivalence removes the most immediate behavioral pathway through which a perturbation can affect the future. For the one-shot feedforward intervention, any later effect must therefore be mediated by the current action, which can be tested by replaying that action. In recurrent policies, however, matching the current action need not determine future behavior because the internal policy state persists across steps.

3.2 Decomposing Perturbation-Induced State Differences

Current action equivalence leaves many possible differences in internal policy state. We therefore ask whether the state difference induced by a real observation perturbation contains structure that is weakly expressed in the current action but becomes more visible later. Let

$$\Delta z_t = z_t^{\text{pert}} - z_t^{\text{clean}} \tag{2}$$

denote the perturbation-induced state difference. Thus, $z_t^{\text{clean}} = z_t^{\text{pert}} - \Delta z_t$. We first separate the part that is locally expressed through the current action. Let

$$J_t = \left.\frac{\partial a_t(z)}{\partial z}\right|_{z=z_t^{\text{pert}}} \in \mathbb{R}^{d_a \times d_z} \tag{3}$$

be the current-action Jacobian at the perturbed state. Its row space contains locally action-visible directions, while its null space gives the local tangent space along which the current action is unchanged to first order. With P_N denoting the orthogonal projector onto $\mathrm{Null}(J_t)$, we decompose

$$\Delta z_t = \Delta z_t^{\text{inst}} + \Delta z_t^{\text{equiv}}, \qquad \Delta z_t^{\text{equiv}} = P_N \Delta z_t, \qquad \Delta z_t^{\text{inst}} = (I - P_N)\Delta z_t. \tag{4}$$

Here, Δz_t^{inst} is the component locally expressed in the current action, whereas $\Delta z_t^{\text{equiv}}$ is locally tangent to current action equivalence. Because this split is local, we later verify exact equality of the actual current action before evaluating future behavior.

We next characterize how directions within this locally action-equivalent component become expressed through future actions. For a fixed future observation sequence, the internal state evolves over time, and the action at step $t + k$ is given by $a_{t+k}(z_{t+k})$. We define

$$B_{t,k} = \left.\frac{\partial a_{t+k}(z_{t+k})}{\partial z_t}\right|_{z_t=z_t^{\text{pert}}}, \tag{5}$$

which measures how a change in the current internal state propagates to the action k steps later. We aggregate these sensitivities over a horizon H using the finite-horizon future-action operator

$$G_{t,H} = \sum_{k=1}^{H} B_{t,k}^\top B_{t,k}, \qquad G_{\text{defer}}(t,H) = P_N G_{t,H} P_N. \tag{6}$$

Here, $G_{\text{defer}}(t,H)$ restricts future-action sensitivity to directions that are locally invisible to the current action. A direction with larger quadratic energy under $G_{\text{defer}}(t,H)$ is weakly expressed through the current action but more visible through actions over the following H steps. Because the observation sequence is fixed, this quantity measures propagation through the policy without mixing in environment dynamics.

Let $\lambda_1 \geq \cdots \geq \lambda_d$ be the eigenvalues of $G_{\text{defer}}(t,H)$. We choose the smallest r such that

$$\frac{\sum_{i=1}^{r} \lambda_i}{\sum_{i=1}^{d} \lambda_i} \geq 0.9,$$

and let P_D project onto the span of the corresponding r eigenvectors. We decompose the locally action-equivalent component as

$$\Delta z_t^{\text{equiv}} = \Delta z_t^{\text{defer}} + \Delta z_t^{\text{rem}}, \qquad \Delta z_t^{\text{defer}} = P_D \Delta z_t^{\text{equiv}}, \qquad \Delta z_t^{\text{rem}} = (I - P_D)\Delta z_t^{\text{equiv}}. \tag{7}$$

We call these the deferred and remaining components, respectively. This decomposition separates components with different temporal action-expression profiles, which we evaluate through controlled interventions in the next section.

3.3 Testing Causal Relevance

We next test whether the deferred and remaining components have different causal effects on future behavior. Let $d_t = \Delta z_t^{\text{defer}}$ and $r_t = \Delta z_t^{\text{rem}}$. Because their magnitudes can differ, we construct equal-norm interventions,

$$\rho_t = \min\{\|d_t\|_2, \|r_t\|_2\}, \qquad \hat{d}_t = \rho_t \frac{d_t}{\|d_t\|_2}, \qquad \hat{r}_t = \rho_t \frac{r_t}{\|r_t\|_2}. \tag{8}$$

Since $\Delta z_t = z_t^{\text{pert}} - z_t^{\text{clean}}$, we intervene from the perturbed state toward the clean state:

$$z_t^{\text{defer}} = z_t^{\text{pert}} - \hat{d}_t, \qquad z_t^{\text{rem}} = z_t^{\text{pert}} - \hat{r}_t. \tag{9}$$

The decomposition is based on local Jacobian geometry, so these interventions need not preserve the current action exactly under the nonlinear policy. Before evaluation, we therefore apply a small finite correction that enforces

$$a_t(z_t^{\text{defer}}) = a_t(z_t^{\text{rem}}) = a_t(z_t^{\text{pert}}). \tag{10}$$

Thus, the deferred and remaining interventions are matched in magnitude and produce exactly the same current action; they differ only in the internal-state direction being changed.

We evaluate their future effects in two complementary settings. Let a_{t+k}^x and s_{t+k}^x denote the action and physical state at step $t+k$ for branch $x \in \{\text{pert}, \text{defer}, \text{rem}, \text{clean}\}$. Under common-observation replay, all branches receive the same recorded future observations, isolating differences that propagate through the policy. We measure cumulative future-action distance to the clean continuation,

$$D_H^{\text{act}}(x) = \sum_{k=1}^{H} \left\| a_{t+k}^x - a_{t+k}^{\text{clean}} \right\|_2 . \tag{11}$$

Under native closed-loop evaluation, branches execute the matched current action and then interact with the environment independently. We measure cumulative standardized trajectory distance,

$$D_H^{\text{traj}}(x) = \sum_{k=1}^{H} \frac{1}{\sqrt{d_s}} \left\| \frac{s_{t+k}^x - s_{t+k}^{\text{clean}}}{\sigma_s + \epsilon} \right\|_2 , \tag{12}$$

where d_s is the state dimension and σ_s contains coordinate-wise scales estimated from the calibration split.

Lower values indicate stronger recovery toward the clean continuation. For reporting, we normalize each intervention by the corresponding perturbed distance,

$$R_H^m(x) = 100 \, \frac{D_H^m(x)}{D_H^m(\text{pert})}, \qquad m \in \{\text{act}, \text{traj}\}, \tag{13}$$

so that 0% denotes complete recovery and 100% denotes no recovery relative to the perturbed branch. A smaller residual for the deferred intervention than for the equal-norm remaining intervention therefore provides evidence that the deferred component is more causally relevant for future behavior.

4 Experiments

4.1 Experimental Setup

Policies and environments. The main recurrent experiments use 12 released LSTM checkpoints across four MuJoCo environments and three policy families. Additional experiments evaluate 60 independently seeded recurrent checkpoints and extend the analysis to Vanilla RNN and GRU architectures. For the tested one-shot feedforward intervention, replaying the current action exactly reproduces the downstream trajectory, with full feedforward results reported in Appendix B.

Perturbations and data splits. The main experiments perturb each observation dimension by a fixed normalized magnitude with an independently sampled positive or negative sign. Additional evaluations use random uniform noise, targeted-action PGD, and maximum-action-deviation PGD. We split evaluation episodes into calibration and test sets. Calibration episodes are used only to estimate statistics and construct the intervention, including observation scales, ranks, and reference bases. All causal outcomes are measured on disjoint test episodes, and test outcomes are never used to construct or select interventions.

Evaluation and inference. We use $H = 10$ as the primary evaluation horizon. For recurrent policies, we evaluate future behavior under two settings. A branch denotes a counterfactual rollout initialized from the same saved environment and policy state but under a different intervention. In

Table 1: Recurrent counterfactuals under current action equivalence. All counterfactuals satisfy exact current action equivalence and share the same first environment transition. Common-observation replay isolates recurrent-state propagation by holding future observations fixed, whereas closed-loop evaluation additionally captures environmental feedback. Deferred directions are constructed using only the next $H{=}10$ steps but remain distinguishable at longer horizons. Entries report mean $\pm$ SD across released checkpoints.

Counterfactual	$H{=}10$ divergence	$H{=}50$ divergence	$H{=}100$ divergence
(a) Common-observation action divergence			
Random cell-only	0.026 ± 0.019	0.029 ± 0.024	0.030 ± 0.024
Random hidden-null	0.034 ± 0.023	0.038 ± 0.028	0.038 ± 0.028
Cell-only direct	0.059 ± 0.049	0.065 ± 0.054	0.065 ± 0.054
Hidden-defer	0.587 ± 0.354	0.631 ± 0.406	0.631 ± 0.406
Full-state defer	$\mathbf{1.098 \pm 0.776}$	$\mathbf{1.178 \pm 0.854}$	$\mathbf{1.179 \pm 0.855}$
(b) Closed-loop trajectory divergence			
Random cell-only	0.840 ± 0.780	17.800 ± 13.290	48.420 ± 34.340
Random hidden-null	0.920 ± 0.780	17.240 ± 12.830	46.950 ± 33.040
Cell-only direct	1.070 ± 0.950	18.220 ± 12.910	48.900 ± 33.680
Hidden-defer	1.870 ± 1.110	21.240 ± 14.120	52.890 ± 34.360
Full-state defer	$\mathbf{2.160 \pm 1.210}$	$\mathbf{22.320 \pm 14.470}$	$\mathbf{54.930 \pm 34.830}$

common observation replay, all branches receive the same recorded future observations, isolating differences that propagate through the policy. In native closed loop evaluation, branches execute the matched current action from the same environment state and then interact with the environment independently. We denote future-action distance under common-observation replay by $D_H^{\mathrm{act}}(x)$ and standardized trajectory distance under native closed-loop evaluation by $D_H^{\mathrm{traj}}(x)$, following Eqs. 11 and 12. Lower values indicate closer recovery toward the clean continuation. When comparing perturbations with different baseline magnitudes, we report the normalized residual $R_H^m(x)$ from Eq. 13. All candidate pairs must satisfy exact current-action and first-transition conditions before future outcomes are evaluated. We aggregate results at the policy or checkpoint level and report mean $\pm$ standard deviation. Detailed eligibility rules and statistical procedures are provided in Appendix A.

4.2 Current Action Equivalence Does Not Imply Future Equivalence

A central question is whether fixing the current action is sufficient to determine future behavior in a recurrent policy. For an LSTM, the internal state consists of the hidden state h_t and cell state c_t, and different internal states can produce the same current action while evolving differently afterward.

Starting from the same environment state, we construct several counterfactuals under current action equivalence. A cell-only perturbation changes c_t while keeping h_t fixed, and therefore leaves the current action unchanged. A hidden-null perturbation changes h_t along a direction in the null space of the current action head while keeping c_t fixed. Although such a direction is invisible to the current action, it can still affect subsequent recurrent updates because h_t enters the LSTM gates. All retained counterfactuals match the raw and clipped current actions and the first environment transition exactly.

Table 1 first evaluates these counterfactuals under common-observation replay, where every branch receives the same future observation sequence. This removes differences caused by environmental feedback and isolates propagation through the internal state. Random cell-only and hidden-null perturbations produce only small future-action divergence, remaining around 0.03–0.04 through $H{=}50$.

Deferred directions show a markedly different temporal profile. The hidden-deferred intervention produces action divergences of 0.59 at $H{=}10$ and 0.63 at $H{=}50$, while the full-state deferred intervention reaches 1.10 and 1.18, respectively. Divergence appears quickly, after roughly 1.4–1.6 recurrent steps across the tested counterfactuals. Crucially, the deferred directions are constructed using only the next $H{=}10$ steps, yet their strong separation from the random controls persists through $H{=}50$. This shows that the identified deferred structure is not confined to the horizon used for its construction.

Table 2: Cross perturbation recurrent comparison across four control environments. The Perturbed column reports $D_{10}^{m}(\mathrm{pert})$; Deferred and Remaining report $R_{10}^{m}(\mathrm{defer})$ and $R_{10}^{m}(\mathrm{rem})$, respectively. Here $m = \mathrm{act}$ for common-observation replay and $m = \mathrm{traj}$ for native closed-loop evaluation. Thus, 0% indicates complete recovery and 100% indicates no recovery. Values are mean $\pm$ SD across 12 released checkpoints.

Observation perturbation	$D_{10}^{m}(\mathrm{pert})$	$R_{10}^{m}(\mathrm{defer})$ (%)	$R_{10}^{m}(\mathrm{rem})$ (%)
Common observations: $m = \mathrm{act}$			
Random sign	0.0311 ± 0.0278	$\mathbf{12.5 \pm 6.9}$	98.4 ± 1.4
Random uniform	0.0106 ± 0.0096	$\mathbf{16.9 \pm 14.1}$	98.1 ± 1.3
Targeted action PGD	0.0603 ± 0.0561	$\mathbf{18.8 \pm 12.9}$	98.1 ± 1.3
Maximum action deviation PGD	0.1266 ± 0.1271	$\mathbf{21.7 \pm 13.8}$	98.3 ± 0.7
Average	0.0571 ± 0.0506	$\mathbf{17.5 \pm 11.3}$	98.2 ± 0.8
Closed loop: $m = \mathrm{traj}$			
Random sign	0.691 ± 0.523	$\mathbf{56.9 \pm 19.2}$	97.6 ± 3.2
Random uniform	0.502 ± 0.357	$\mathbf{55.1 \pm 21.6}$	100.0 ± 7.1
Targeted action PGD	0.903 ± 0.797	$\mathbf{55.5 \pm 21.7}$	99.0 ± 2.5
Maximum action deviation PGD	0.978 ± 0.693	$\mathbf{58.6 \pm 20.9}$	100.3 ± 2.4
Average	0.768 ± 0.582	$\mathbf{56.5 \pm 18.9}$	99.2 ± 2.9

The same pattern appears under native closed-loop execution. At $H{=}10$, trajectory divergence is 0.84 and 0.92 for the two random controls, compared with 1.87 for the hidden-deferred intervention and 2.16 for the full-state deferred intervention. The separation remains visible over much longer rollouts, with the deferred interventions reaching 52.89 and 54.93 at $H{=}100$.

Together, these results show that current action equivalence does not imply future equivalence in recurrent policies. More importantly, the future effect is not generic to arbitrary state changes under current action equivalence: deferred directions produce substantially larger later behavioral differences than random cell-only or hidden-null perturbations. The separation remains evident when evaluated beyond the $H = 10$ construction horizon, suggesting that the identified structure is not specific to the horizon used to define it. The next section asks whether real observation perturbations actually populate current-action-silent directions, whether these directions are preferentially aligned with future-sensitive geometry, and whether this structure is causally relevant.

4.3 Deferred structure and causal relevance

Do real perturbations induce internal state differences that are weakly expressed in the current action but become relevant later? Across the 12 evaluated LSTM checkpoints, 7.27% of the perturbation-induced state displacement energy lies in the locally action-visible component, whereas 92.73% lies in the locally action-equivalent component. This large fraction is partly expected from the higher dimensionality of the recurrent state relative to the action output. More importantly, the action-equivalent component is not occupied uniformly: real perturbations are preferentially aligned with future-sensitive directions compared with norm-matched random directions in the same null space. The deferred projection accounts for 11.38% of the action-equivalent displacement energy, suggesting that the relevant structure is concentrated along a subset of future-sensitive directions rather than spread broadly across the null space.

The same equal-norm comparison is further evaluated across 60 recurrent checkpoints spanning four environments, three training families, and five seed lineages per configuration. The deferred advantage remains positive across all environments, with mean raw-distance contrasts $D_{10}^{\mathrm{act}}(\mathrm{rem}) - D_{10}^{\mathrm{act}}(\mathrm{defer}) = 0.108$ under common observations and $D_{10}^{\mathrm{traj}}(\mathrm{rem}) - D_{10}^{\mathrm{traj}}(\mathrm{defer}) = 0.348$ for closed-loop trajectories. Full results are reported in Appendix C.9.

Table 2 tests whether the deferred structure is causally relevant across different observation perturbations. We compare equal norm interventions along the deferred and remaining components while preserving the current action and the first physical transition.

Under common observations, $R_{10}^{\mathrm{act}}(\mathrm{defer})$ is 17.5% on average, whereas $R_{10}^{\mathrm{act}}(\mathrm{rem})$ is 98.3%. This separation is consistent across random sign, random uniform, targeted action PGD, and maximum

Table 3: Directional specificity within the same deferred subspace. The intervention branches start from the same perturbed state and use the same target P_D, intervention norm, current action, and first physical transition, differing only in the intervention direction within P_D. Lower distances indicate closer recovery toward the clean continuation. Entries are means $\pm$ standard deviations across 12 released checkpoints.

Intervention	Future action	Closed-loop trajectory	Reward error
Distance to clean continuation (lower is better)			
Real deferred intervention	0.0257 ± 0.0349	0.6753 ± 0.5651	0.1461 ± 0.1716
No intervention (attacked)	0.1244 ± 0.1193	1.0792 ± 0.7980	0.2877 ± 0.2401
Random direction in the same P_D	0.1810 ± 0.1530	1.1199 ± 0.7576	0.3432 ± 0.2746
Opposite of the real deferred direction	0.3718 ± 0.3266	1.3773 ± 0.7983	0.4738 ± 0.3707
Difference from the real deferred intervention (positive means real is better)			
Random in P_D − Real	0.1553 ± 0.1392	0.4446 ± 0.2237	0.1971 ± 0.1185
Opposite − Real	0.3461 ± 0.3178	0.7020 ± 0.3110	0.3277 ± 0.2013

action deviation PGD, with deferred residual distances ranging from 12.5% to 21.7% and remaining residual distances staying near 98%.

The same ordering persists in native closed-loop evaluation: $R_{10}^{\mathrm{traj}}(\mathrm{defer}) = 56.5\%$ on average, compared with $R_{10}^{\mathrm{traj}}(\mathrm{rem}) = 99.2\%$. Thus, across qualitatively different observation perturbations, changing the deferred component consistently moves future behavior toward the clean continuation, whereas an equal norm change in the remaining component produces little recovery. Because the two interventions preserve the initiating action and first physical transition, this difference reflects the internal state direction being changed rather than the intervention magnitude or the initiating action.

The gap between the two settings reflects what each controls. Under common-observation replay, future observations are held fixed, so divergence traces purely to the internal state at the intervention point; in native closed-loop evaluation, residual state differences propagate through environment feedback across subsequent steps. The deferred advantage is substantial in both settings, confirming that the effect is not an artifact of the controlled replay.

Task-level reward corroborates these findings. Restoring the deferred component reduces reward error to 0.146 relative to the clean continuation, compared with 0.288 for the unmodified perturbed state—a reduction of approximately 49%. An equal-norm restoration of the remaining component yields no comparable improvement. This reward advantage holds across all tested perturbation types. Under a receding adaptive-release protocol that gradually steps down the intervention magnitude over subsequent steps, the same ordering persists across all tested amplitudes and release schedules (Appendix C.7).

4.4 Directional Specificity Within the Deferred Subspace

Directional specificity is examined by comparing the real deferred intervention with equal-norm interventions along random directions sampled from the same target P_D. For each branch, the random-direction control shares the same perturbed state, deferred subspace, intervention norm, current action, and first physical transition, differing only in the direction of intervention within P_D. An opposite-direction control is also constructed by moving along the real deferred axis in the reverse direction with the same intervention norm.

Table 3 shows a clear directional ordering. Under common-observation replay, the real deferred intervention produces the smallest distance to the clean continuation. Leaving the attacked state unchanged results in a larger distance, while an equal-norm random direction within the same P_D moves the future actions still farther from the clean continuation. Moving in the opposite direction produces the largest deviation. The same ordering is observed under native closed-loop evaluation.

At all checkpoints, the four conditions exhibited a consistent magnitude relationship: real deferred was the smallest, attacked was the second smallest, random in P_D was larger, and opposite direction was the largest. Thus, selecting a direction from the deferred subspace alone does not determine the recovery effect. The direction induced by the real perturbation produces substantially stronger

Table 4: Deferred recovery across recurrent architectures. Results cover LSTM, Vanilla RNN, and GRU policies across four environments. The table reports $D^m_{10}(\mathrm{pert})$, $R^m_{10}(\mathrm{defer})$, and $R^m_{10}(\mathrm{rem})$ using the notation of Section 3.3. Relative reduction is $100\,[1 - R^m_{10}(\mathrm{defer})/R^m_{10}(\mathrm{rem})]\%$. Results are averaged equally over four observation perturbations and reported as mean $\pm$ SD across checkpoint settings. All interventions have equal norm and satisfy exact current action equivalence.

Architecture	$D^m_{10}(\mathrm{pert})$	$R^m_{10}(\mathrm{defer})$ (%)	$R^m_{10}(\mathrm{rem})$ (%)	Relative reduction (%)
(a) Common observations: $m = \mathrm{act}$				
LSTM	0.057 ± 0.051	17.5 ± 11.3	98.2 ± 0.8	82.2 ± 11.6
Vanilla RNN	0.036 ± 0.015	22.0 ± 10.7	97.3 ± 1.2	77.4 ± 10.8
GRU	0.034 ± 0.025	24.8 ± 10.1	97.9 ± 1.3	74.7 ± 10.3
(b) Closed loop: $m = \mathrm{traj}$				
LSTM	0.768 ± 0.582	56.5 ± 18.9	99.2 ± 2.9	43.2 ± 18.2
Vanilla RNN	0.736 ± 0.670	55.3 ± 16.9	99.3 ± 2.0	44.3 ± 16.7
GRU	0.845 ± 0.634	58.3 ± 12.7	99.2 ± 2.6	41.2 ± 11.5

recovery than an equal-norm random direction in the same subspace, while reversing that direction further increases the deviation from the clean continuation.

The same directional pattern appears at the task-reward level under the receding adaptive-release protocol. The real deferred intervention produces the smallest clean-relative reward error, followed by the attacked continuation, the random direction in P_D, and the opposite-direction intervention. The random-minus-real reward difference is positive across all 12 checkpoints, showing that the directional dependence observed in future actions and closed-loop trajectories is also reflected in task return.

Together, these results show that recovery depends not only on the deferred subspace itself, but also on the specific perturbation-induced direction within that subspace.

4.5 Transfer Across Recurrent Architectures

The main analysis uses LSTM policies. To test whether the deferred advantage depends on this recurrent architecture, Table 4 repeats the paired intervention with Vanilla RNN and GRU policies across all four environments and four observation perturbations. The same ordering appears across all three architectures. Under common observations, $R^{\mathrm{act}}_{10}(\mathrm{defer})$ is 17.5%, 22.0%, and 24.8% for LSTM, Vanilla RNN, and GRU, respectively, whereas $R^{\mathrm{act}}_{10}(\mathrm{rem})$ remains approximately 97–98%. The corresponding relative reductions are 82.2%, 77.4%, and 74.7%. The ordering remains consistent in native closed loop evaluation: $R^{\mathrm{traj}}_{10}(\mathrm{defer})$ is 55–58%, compared with approximately 99% for $R^{\mathrm{traj}}_{10}(\mathrm{rem})$, giving relative reductions of 41.2–44.3%. The deferred advantage therefore extends beyond LSTM policies and remains consistent across the three recurrent architectures tested here.

The deferred intervention shows a consistent advantage over the remaining component intervention across different future observation streams, construction horizons, trace cutoffs, donor geometries, and recurrent-state representations. Full results are reported in Appendix C.5.

5 Conclusion

Current action equivalence separates effects already expressed through the current action from effects that remain active inside the policy. For the tested feedforward intervention, the downstream effect is fully explained by the initiating action. In recurrent policies, real observation perturbations can additionally induce deferred internal state components that preserve the current action yet causally influence later behavior. This distinction is consistent across perturbations, recurrent architectures, environments, and independently trained checkpoints, providing a way to identify when perturbation-induced internal state becomes behaviorally relevant.

Reproducibility Statement

All statistics use frozen policy or checkpoint aggregates. Appendix A documents branch construction, data separation, eligibility criteria, perturbations, and bootstrap procedures, while Appendix E reports the software and hardware environment.

AI Use Statement

Generative AI tools assisted with code development, manuscript editing, and data analysis. All reported results and conclusions were reviewed and verified by the authors, who remain fully responsible for the work.

References

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. *Advances in neural information processing systems*, 34:9574–9586, 2021.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Causal Learning and Reasoning*, pp. 160–187. PMLR, 2024.

Bear Häon, Kaylene Stocking, Ian Chuang, and Claire Tomlin. Mechanistic interpretability for steering vision-language-action models. *arXiv preprint arXiv:2509.00328*, 2025.

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*, 2024.

Sandy Huang, Nicolas Papernot, Ian Goodfellow, Yan Duan, and Pieter Abbeel. Adversarial attacks on neural network policies. *arXiv preprint arXiv:1702.02284*, 2017.

Chonghe Jiang, Bao Nguyen, Anthony Man-Cho So, and Viet Anh Nguyen. Probe-free low-rank activation intervention. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2812–2824, 2025.

Yu Kawano and Jacquelien MA Scherpen. Empirical differential gramians for nonlinear model reduction. *Automatica*, 127:109534, 2021.

Mohamad H Kazma and Ahmad F Taha. Observability for nonlinear systems: connecting variational dynamics, lyapunov exponents, and empirical gramians. *IEEE Transactions on Control of Network Systems*, 2026.

Armin Kekić, Jan Schneider, Dieter Büchler, Bernhard Schölkopf, et al. Learning nonlinear causal reductions to explain reinforcement learning policies. In *International Conference on Learning Representations*, volume 2026, pp. 118426–118455, 2026.

Wenhao Lu, Xufeng Zhao, Thilo Fryen, Jae Hee Lee, Mengdi Li, Sven Magg, and Stefan Wermter. Causal state distillation for explainable reinforcement learning. In *Causal Learning and Reasoning*, pp. 106–142. PMLR, 2024.

Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. Explainable reinforcement learning through a causal lens. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 2493–2500, 2020.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. In *International Conference on Learning Representations*, volume 2024, pp. 26486–26515, 2024.

Trevor A McInroe, Michael Spurrier, Jennifer Sieber, and Stephen Conneely. Analyzing the hidden activations of deep policy networks: Why representation matters. *arXiv preprint arXiv:2103.06398*, 2021.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.

Ulisse Mini, Peli Grietzer, Mrinank Sharma, Austin Meek, Monte MacDiarmid, and Alexander Matt Turner. Understanding and controlling a maze-solving policy network. *arXiv preprint arXiv:2310.08043*, 2023.

Anay Pattanaik, Zhenyi Tang, Shuijing Liu, Gautham Bommannan, and Girish Chowdhary. Robust deep reinforcement learning with adversarial attacks. *arXiv preprint arXiv:1712.03632*, 2017.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

Wei Tang, Xinyan Jiang, Fakhri Karray, and Lijie Hu. In-context learning operates as concept subspace learning. *arXiv preprint arXiv:2605.18830*, 2026.

Mohammad Taufeeque, Aaron Tucker, Adam Gleave, and Adrià Garriga-Alonso. Path channels and plan extension kernels: A mechanistic description of planning in a sokoban rnn. In *International Conference on Learning Representations*, volume 2026, pp. 134244–134277, 2026.

Tristan Trim and Triston Grayston. Mechanistic interpretability of reinforcement learning agents. *arXiv preprint arXiv:2411.00867*, 2024.

Sankaran Vaidyanathan, David Arbour, Aaron Mueller, Scott Niekum, and David Jensen. The curse of multiple mediators: Hidden interaction effects in activation patching. *arXiv preprint arXiv:2606.27510*, 2026.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, volume 2024, pp. 1651–1678, 2024.

Huan Zhang, Hongge Chen, Chaowei Xiao, Bo Li, Mingyan Liu, Duane Boning, and Cho-Jui Hsieh. Robust deep reinforcement learning against adversarial perturbations on state observations. *Advances in neural information processing systems*, 33:21024–21037, 2020.

Jiapeng Zhu, Ruili Feng, Yujun Shen, Deli Zhao, Zheng-Jun Zha, Jingren Zhou, and Qifeng Chen. Low-rank subspaces in gans. *Advances in Neural Information Processing Systems*, 34:16648–16658, 2021.

A Experimental Protocol

A.1 Branch construction and data separation

Each branch starts from a serialized simulator snapshot collected during a clean episode. The snapshot fixes the physical state and episode time. Clean, perturbed, candidate, and action-replay branches use the same frozen policy and matched deterministic seeds. A candidate recurrent intervention replaces one internal state at the snapshot; its initiating action is applied to a clone of the snapshot, after which the frozen policy runs without further state intervention.

Evaluation episodes are split into calibration and test sets. Calibration episodes are used only to estimate observation scales and any ranks or reference bases fixed before test evaluation. All causal outcomes are measured on disjoint test episodes, and test outcomes are not used to construct or select an intervention. The source perturbation is a branch-specific random-sign observation perturbation. Held-out perturbations are random-uniform noise, targeted-action PGD, and maximum-action-deviation PGD.

The primary horizon is $H = 10$; robustness checks construct the operator at $H = 5$ or evaluate at $H = 20$, while current-action-equivalent boundary checks also report $H = 50$ and $H = 100$. Horizon distances use the notation of Section 3.3: D_H^{act} is evaluated under common-observation replay, D_H^{traj} under native closed-loop execution, and R_H^m is the corresponding residual normalized by the perturbed branch. Lower distances indicate closer recovery to the clean continuation.

A.2 Eligibility and inference

Before any future-action, trajectory, return, or recovery outcome is computed, each candidate pair is required to satisfy exact raw and clipped current-action equality and the same first physical transition. Equal-norm comparisons also require nonzero interventions and the designated relative norm tolerance. Branches that fail these pre-outcome checks are excluded only from the corresponding contrast. The 12 released recurrent checkpoints use evaluation seeds 2000–2023 and contain 432 attempted branches; the independent replication uses 2,160 attempted branches across 60 checkpoints.

We first aggregate within each trained policy or checkpoint, then report the mean $\pm$ standard deviation over those inference units. Headline intervals use 5,000 paired bootstrap draws with seed 42, resampling policies or distinct checkpoints while retaining branch-level rows and replicas within each unit. The 60-checkpoint replication uses an environment-to-seed hierarchical bootstrap. A directional contrast is supported when its 95% interval lies strictly in the hypothesized direction.

B Feed-forward Boundary Check

The feed-forward analysis is retained only as a boundary condition for the recurrent study. We replay the initiating action produced by the tested one-shot intervention while leaving the hidden state unchanged, and then run the deterministic downstream rollout. The maximum restored-action replay discrepancy is exactly 0 across all 972 replays. Thus, for this tested feed-forward intervention, reproducing the initiating action reproduces the downstream trajectory. The recurrent experiments below test the complementary case in which exact current-action and first-transition matching still leaves the internal policy state free to affect later actions.

C Recurrent Current-action Equivalence and Deferred Effects

C.1 Current-action-equivalent recurrent counterfactuals

The recurrent boundary uses 12 distinct released LSTM checkpoints over Ant, HalfCheetah, Hopper, and Walker2d, with one checkpoint for each of the three training families LSTM-PPO, ATLA-LSTM-PPO, and ATLA-LSTM-SAPPO, as summarized in Table 5.

Table 5: Released recurrent checkpoint population. Every cell represents one distinct checkpoint; release metadata omit training seeds and parent-run lineage.

Environment	LSTM-PPO	ATLA-LSTM-PPO	ATLA-LSTM-SAPPO	Total
Ant-v4	1	1	1	3
HalfCheetah-v4	1	1	1	3
Hopper-v4	1	1	1	3
Walker2d-v4	1	1	1	3
Total	4	4	4	12

C.2 Real perturbation-induced decomposition

For each clean–perturbed pair, let $\Delta z_t = z_t^{\mathrm{pert}} - z_t^{\mathrm{clean}}$. At the perturbed state, the current-action Jacobian J_t defines the locally action-visible row space and its orthogonal null space. With P_N the projector onto $\mathrm{Null}(J_t)$, the displacement is split as $\Delta z_t = \Delta z_t^{\mathrm{inst}} + \Delta z_t^{\mathrm{equiv}}$, where $\Delta z_t^{\mathrm{equiv}} = P_N \Delta z_t$. Within this current-action-equivalent component, the finite-horizon operator $G_{\mathrm{defer}}(t, H) = P_N(\sum_{k=1}^{H} B_{t,k}^{\top} B_{t,k}) P_N$ measures future-action expression under fixed observations. The deferred subspace is the smallest leading eigenspace retaining 90% of the operator trace, and $\Delta z_t^{\mathrm{equiv}} = \Delta z_t^{\mathrm{defer}} + \Delta z_t^{\mathrm{rem}}$ uses the orthogonal remainder. The remaining component is low visibility under this finite-horizon operator, not necessarily zero in future behavior.

The reconstruction error is below 10^{-8}. The real paired null component has Rayleigh score 0.071 (0.040–0.106), compared with 0.030 (0.017–0.046) for an independently seeded norm-matched random-null control; the paired-minus-random contrast is 0.041 (0.018–0.066). Thus, real perturbations contain deferred structure while retaining a non-deferred remainder.

C.3 Equal-norm causal comparison

The deferred and remaining interventions are two directions extracted from the same perturbation-induced displacement and start from the same perturbed recurrent state. Equal-norm matching removes intervention magnitude as an explanation. A small deterministic correction restores the exact current action; only branches that also match the first physical transition are evaluated. Of 432 branches, 411 satisfy the source-action, equal-norm, and transition checks. Before matching, the remaining norm is 1.786 (1.433–2.140) and the deferred norm is 0.495 (0.404–0.589); their fractions of the full current-null component are 0.951 (0.939–0.963) and 0.282 (0.242–0.324), respectively. Table 6 gives the resulting causal comparison.

Table 6: Branch-wise equal-norm paired decomposition at $H = 10$. Common- observation entries are D_{10}^{act} to the clean recurrent reference; the closed-loop entry is the remaining-minus-deferred contrast in D_{10}^{traj}. Descriptive distances report means only; parentheses report 95% bootstrap confidence intervals for inferential contrasts over 12 checkpoints and 411 valid branches.

Quantity	Estimate
Attacked distance	0.127
Deferred equal-norm distance	0.031
Remaining equal-norm distance	0.126
Remaining − deferred	0.095 (0.050–0.151)
Closed-loop trajectory contrast	0.315 (0.229–0.404)

The environment-to-checkpoint hierarchical bootstrap gives the same primary common-observation contrast, 0.095 (0.037–0.170). Because both interventions preserve the initiating action and first transition, their different future outcomes identify the changed recurrent direction rather than the current action or intervention norm.

C.4 COMMON-OBSERVATION AND NATIVE CLOSED-LOOP EVALUATION

Common-observation replay feeds every branch the same recorded future observation sequence. It therefore isolates propagation through recurrent policy state and is summarized by D_H^{act} or R_H^{act}. Native closed-loop evaluation applies the matched initiating action once and then lets each branch interact with the environment independently; it measures the resulting physical consequence through D_H^{traj} or R_H^{traj}. The equal-norm comparison gives a positive deferred advantage in both settings, with the closed-loop trajectory contrast 0.315 (0.229–0.404).

C.5 ROBUSTNESS TO CONSTRUCTION CHOICES

The following checks change one potential alternative explanation at a time and retain the same exact-action and first-transition gates. The cross-perturbation experiment covers random sign, random uniform, targeted-action PGD, and maximum-action-deviation PGD; the deferred-over-remaining ordering persists in both R_{10}^{act} and R_{10}^{traj}. Across recurrent architectures, the same ordering holds for LSTM, Vanilla RNN, and GRU policies: common-observation deferred residuals are 17.5%, 22.0%, and 24.8%, versus approximately 97–98% for remaining, with relative reductions of 82.2%, 77.4%, and 74.7%.

C.5.1 CROSS-HORIZON CONFIRMATION

We construct $G_{\mathrm{defer}}(t, 5)$ using only the next five actions, form its 90%-trace subspace, and evaluate the resulting equal-norm interventions at $H = 10$ and $H = 20$. Exact-action checks retain 405 of 432 branches with an identical first physical transition. Table 7 reports the common-observation and native closed-loop contrasts.

Table 7: Cross-horizon validation of an $H{=}5$-constructed paired decomposition. Distances are to the clean recurrent reference; the contrast is $D_H^{\mathrm{act}}(\mathrm{rem}) - D_H^{\mathrm{act}}(\mathrm{defer})$ except for the closed-loop row. Deferred and Remaining report means; parentheses report 95% bootstrap confidence intervals for contrasts over 12 checkpoints and 405 branches.

Evaluation	Deferred	Remaining	Contrast
Common observations, $H{=}10$	0.035	0.126	0.091 (0.048–0.143)
Common observations, $H{=}20$	0.038	0.134	0.096 (0.050–0.152)
Native closed-loop trajectory, $H{=}20$			0.827 (0.526–1.179)

C.5.2 ATTACKED-OBSERVATION CONSTRUCTION

Recomputing $G_{\mathrm{defer}}(t, 10)$ along the attacked future observation stream, while evaluating common-observation outcomes on the clean stream, retains 411 valid branches and the same ordering, as shown in Table 8.

Table 8: Paired equal-norm results when temporal geometry is constructed along the attacked future observation stream. Common-observation distances are to the clean recurrent reference. Descriptive distances report means; parentheses report 95% bootstrap confidence intervals for contrasts over 12 checkpoints and 411 valid branches.

Quantity	Estimate
Deferred clean distance	0.036
Remaining clean distance	0.125
Remaining − deferred	0.089 (0.043–0.150)
Closed-loop trajectory contrast	0.315 (0.231–0.401)

C.5.3 EPISODE-SHUFFLED TEMPORAL GEOMETRY

For each target branch, a same-checkpoint donor from a different episode supplies the temporal operator, while the target displacement, observation stream, and rollout remain fixed. All resulting

components share the smallest branch-level norm, and exact current-action and first-transition checks retain 402 branches. Table 9 compares the target and donor geometries.

Table 9: Target versus episode-shuffled temporal geometry. Positive values favor deferred over remaining, except the final row, which compares target and donor deferred components. Parentheses report 95% bootstrap confidence intervals over 12 checkpoints.

Contrast	Common observations	Native closed-loop trajectory
Target geometry	0.087 (0.047–0.134)	0.262 (0.189–0.339)
Donor geometry	0.014 (0.005–0.025)	0.040 (0.016–0.064)
Target − donor specificity	0.073 (0.039–0.113)	0.222 (0.152–0.297)
Target deferred over donor deferred	0.050 (0.025–0.080)	0.162 (0.106–0.223)

The donor operator retains a small ordering expected from generic future sensitivity, but the target-minus-donor specificity contrast is positive under both evaluations. The effect therefore depends on target-branch temporal geometry rather than on any future-sensitive subspace.

C.5.4 Trace-cutoff sensitivity

Changing the retained trace mass from 80% to 95% changes the mean deferred dimension but not the equal-norm ordering (Table 10).

Table 10: Trace-cutoff sensitivity of the $H{=}10$ equal-norm paired recurrent contrast. The outcome is $D_{10}^{\mathrm{act}}(\mathrm{rem}) - D_{10}^{\mathrm{act}}(\mathrm{defer})$. Parentheses report 95% bootstrap confidence intervals over 12 checkpoints.

Trace cutoff	Branches	Mean deferred dimension	Contrast
80%	412	4.48	0.090 (0.043–0.148)
90%	409	6.45	0.096 (0.048–0.158)
95%	408	8.39	0.099 (0.050–0.162)

C.5.5 Direct observation-space construction

The same distinction can be constructed directly in normalized observation space, without first decomposing a hidden displacement. Both perturbations use one common ℓ_2 norm within the normalized $\ell_\infty = 0.15$ box. A branch is retained only when raw and clipped actions, target norm, recurrent-state change, and first transition satisfy the exact pre-outcome gates. This retains 288 of 432 branches, with 13–34 per checkpoint. Table 11 reports the resulting action and trajectory contrasts.

Table 11: Direct observation-space pairs under exact current-action equivalence. Entries are $H{=}10$ distances from the clean continuation; the contrast is deferred minus remaining in the displayed outcome. Parentheses report 95% bootstrap confidence intervals for the contrast over 12 checkpoints.

Outcome	Deferred	Remaining	Contrast
Future action	4.47×10^{-5}	3.13×10^{-10}	4.47×10^{-5} (2.09×10^{-5}–7.39×10^{-5})
Closed-loop trajectory	0.089	3.00×10^{-4}	0.089 (0.036–0.148)

Both contrasts are positive for all 12 checkpoint means, showing that the temporal distinction is reproducible at the observation interface as well as through an internal-state decomposition.

C.5.6 Function-preserving representation changes

We recompute the complete construction after function-preserving diagonal and dense changes of recurrent hidden coordinates. The diagonal family scales half the coordinates by 1 and half by $\kappa \in \{1, 2, 4, 8\}$. The dense family uses a deterministic dense orthogonal matrix with singular values geometrically spaced from 1 to κ; the native policy is wrapped with the inverse map. In float32, the dense audit has maximum retained raw-sequence error 9.78×10^{-6} and maximum state round-trip

error 3.16×10^{-6}. Dense retained populations are 408, 408, 409, and 407 branches at $\kappa = 1, 2, 4, 8$. Table 12 reports the recomputed contrasts.

Table 12: Function-preserving recurrent representation robustness. Entries are $H{=}10$ equal-norm $D_{10}^{\mathrm{act}}(\mathrm{rem}) - D_{10}^{\mathrm{act}}(\mathrm{defer})$ contrasts over 12 checkpoints; geometry is recomputed at every κ. Parentheses report 95% bootstrap confidence intervals.

κ	Recurrent diagonal	Recurrent dense
1	0.095 (0.050–0.151)	0.095 (0.049–0.150)
2	0.094 (0.049–0.151)	0.093 (0.049–0.145)
4	0.093 (0.049–0.147)	0.084 (0.044–0.138)
8	0.086 (0.046–0.138)	0.090 (0.047–0.145)

The deferred-over-remaining ordering remains positive across both tested function-preserving families, including dense mixing of all hidden coordinates.

C.6 Finite-horizon operator sanity check

The spectrum check is included only to validate the temporal operator used to define the deferred subspace. From $H{=}1$ to $H{=}10$, its trace and effective rank increase while the leading eigenvalue tracks the growth of future-action sensitivity (Table 13).

Table 13: Temporal spectrum within the local tangent space of current-action equivalence. Entries are checkpoint means over 12 checkpoints.

Horizon	Trace	Effective rank	Top eigenvalue
$H{=}1$	2.080	3.427	1.051
$H{=}2$	2.797	3.967	1.300
$H{=}5$	3.651	4.419	1.619
$H{=}10$	4.320	4.578	1.889

As a matched-objective nonlinear sanity check, equal-norm leading and remaining directions have $H{=}10$ future-action energies of 4.998 and 0.206, respectively, and native trajectory divergences of 1.850 and 0.172 (Table 14).

Table 14: Sanity check using matched deferred and remaining directions at $H{=}10$. Entries are checkpoint means. The table validates the temporal operator; it is not an additional causal estimand.

Outcome	Deferred	Remaining	Deferred − remaining
Future-action energy	4.998	0.206	4.792
Trajectory divergence	1.850	0.172	1.678

C.7 Reward-Level Validation

The causal effect of the deferred component is further evaluated at the task-reward level. The results are shown in Table 15. Reward is measured under a receding adaptive-release protocol, where the recurrent-state decomposition is recomputed during the rollout and the deferred or remaining component is restored accordingly. The evaluation uses reward horizon $H = 10$, matching the primary horizon used throughout the main experiments.

For each branch, the reward advantage is defined as

$$A_{10}^{\mathrm{reward}} = \frac{|R_{\mathrm{rem}} - R_{\mathrm{clean}}| - |R_{\mathrm{defer}} - R_{\mathrm{clean}}|}{|R_{\mathrm{att}} - R_{\mathrm{clean}}|},$$

where a positive value indicates that restoring the deferred component moves the return closer to the clean continuation than restoring the remaining component.

Table 15: Reward advantage of the deferred intervention at reward horizon $H = 10$ under the receding adaptive-release protocol. Rows vary the intervention amplitude and columns vary the number of release steps K. Positive values indicate that the deferred intervention recovers the clean return more effectively than the remaining-component intervention. Entries are mean $\pm$ SD across released checkpoints.

Amplitude	$K = 1$	$K = 2$	$K = 3$	$K = 5$	$K = 7$	$K = 10$
0.5	0.276 ± 0.184	0.379 ± 0.166	0.414 ± 0.188	0.421 ± 0.295	0.495 ± 0.318	0.562 ± 0.203
1.0	0.494 ± 0.215	0.557 ± 0.179	0.523 ± 0.172	0.447 ± 0.176	0.546 ± 0.397	0.503 ± 0.258
1.5	0.378 ± 0.216	0.452 ± 0.194	0.450 ± 0.201	0.529 ± 0.176	0.503 ± 0.297	0.462 ± 0.270
2.0	0.180 ± 0.223	0.144 ± 0.157	0.056 ± 0.470	0.202 ± 0.390	0.213 ± 0.676	0.201 ± 0.407

Table 16: Sensitivity of the reward-level comparison to the geometry horizon. Entries are the same quantities as in the reward column of Table 3, recomputed with geometry horizon 20. Lower error is closer to the clean return. Arm entries are mean $\pm$ SD across 12 released checkpoints.

	H_{select}=10	H_{select}=20
Reward error to clean (lower is better)		
Real deferred intervention	0.1461 ± 0.1716	0.1475 ± 0.1705
No intervention (attacked)	0.2877 ± 0.2401	0.2877 ± 0.2401
Random direction in the same P_D	0.3432 ± 0.2746	0.3400 ± 0.2618
Opposite of the real deferred direction	0.4738 ± 0.3707	0.4742 ± 0.3708
Remaining (published baseline)	0.3005 ± 0.2608	0.3024 ± 0.2573
Difference from the real deferred intervention		
Random in P_D − real	0.1971 ± 01185	0.1925 ± 0.1183
Opposite − real	0.3277 ± 0.2013	0.3267 ± 0.2307

The reward column of Table 3 in the main text uses geometry horizon $H_{\text{select}} = 10$, consistent with the deferred-subspace convention used throughout this paper, together with reward horizon $H = 10$. The geometry horizon 20 arm is retained here as a sensitivity analysis (Table 16).

C.8 Finite-Correction and Direction-Preservation Audit

The deferred-subspace construction is based on local Jacobian geometry, while the final interventions are required to match the current action under the nonlinear policy. A finite correction is therefore applied before evaluation. To verify that this correction does not materially change the intended intervention, the correction magnitude and final direction are audited for the within-P_D control.

As shown in Table 17, across the 396 retained branches, the correction is negligible relative to the intervention itself: the maximum correction-to-intervention norm ratio is 9.6×10^{-6}. The cosine similarity between the intended and corrected intervention directions remains 1.0 to numerical precision. The sampled random controls also remain within the target deferred subspace, with a maximum relative projection error of 3.68×10^{-15}. The final current-action residual is at most 1.9×10^{-6}, and all retained branches satisfy the clipped-action and first-transition checks.

These audits show that the comparison in Section 4.4 is not driven by the finite current-action correction: the correction is several orders of magnitude smaller than the intervention and leaves its direction effectively unchanged.

C.9 Independent-lineage recurrent replication

The replication matrix crosses Ant, HalfCheetah, Hopper, and Walker2d with ATLA-LSTM-PPO, ATLA-LSTM-SAPPO, and LSTM-PPO, using five seed lineages in each cell. All 60 checkpoints enter the environment-to-seed hierarchical bootstrap. The exact-action and first-transition gate retains 2,077 of 2,160 attempted branches (96.2%). Positive values below denote lower distance for the equal-norm deferred component than for the equal-norm remaining component. Table 18 reports the overall replication.

Table 17: Audit of the finite current-action correction used in the within-P_D directional control. Projection error is $\|P_D u - u\|_2 / \|u\|_2$ for sampled random directions.

Audit quantity	Result
Retained branches	396
Valid random-P_D directions	12,672
Random-direction acceptance rate	82.8%
Maximum correction / intervention norm	9.6×10^{-6}
Minimum intended–final cosine similarity	1.0
Maximum P_D projection error	3.68×10^{-15}
Maximum current-action residual	1.9×10^{-6}
Clipped-action mismatches	0
First-transition mismatches	0

Table 18: Overall 60-checkpoint recurrent replication. Distances are to the paired clean continuation. Entries are hierarchical means; parentheses report 95% bootstrap confidence intervals.

Evaluation	Deferred	Remaining	Improvement
Common observations	0.021 (0.014–0.030)	0.129 (0.069–0.193)	0.108 (0.056–0.161)
Closed-loop trajectory	0.652 (0.231–1.112)	1.000 (0.458–1.563)	0.348 (0.217–0.475)
Closed-loop action	0.861 (0.350–1.459)	1.650 (0.705–2.863)	0.789 (0.342–1.488)

The common-observation and trajectory contrasts are positive for every checkpoint; the action contrast is positive for 95%. All leave-one- environment-out common-observation and trajectory intervals have positive lower bounds, and the ordering appears in each training family, as detailed in Table 19.

Table 19: Replication by recurrent policy family. Columns report the remaining-minus-deferred improvement; parentheses report 95% hierarchical bootstrap confidence intervals.

Policy family	Common observations	Closed-loop trajectory
ATLA-LSTM-PPO	0.169 (0.091–0.246)	0.384 (0.207–0.529)
ATLA-LSTM-SAPPO	0.045 (0.014–0.095)	0.285 (0.132–0.424)
LSTM-PPO	0.110 (0.059–0.190)	0.374 (0.249–0.504)

Ten ATLA–Ant policies required a numerical training repair and resumed from the latest finite checkpoint in their own seed lineages. Removing all ten repaired policies leaves 50 checkpoints; the hierarchical common-observation, closed-loop trajectory, and closed-loop action contrasts remain positive at 0.111 (0.063–0.162), 0.358 (0.215–0.494), and 0.808 (0.365–1.479), respectively.

D LIMITATIONS

The decomposition is a paired diagnostic that uses a clean reference state and local derivatives. It is not a unique latent decomposition or a formal natural indirect-effect estimator, and it does not imply that all unused pathways are irrelevant. Exact current-action matching removes the initiating-action explanation for the reported branches, but interactions among mediators can remain outside the construction.

The deferred subspace depends on finite-horizon local geometry and on the chosen trace cutoff. Stability across 80%–95% cutoffs, construction at $H=5$ with evaluation at longer horizons, and the attacked-observation check support the reported ordering without establishing a unique temporal basis. Likewise, the representation study covers only the tested diagonal and dense function-preserving transformations with condition numbers at most 8.

The evidence is limited to the evaluated MuJoCo control tasks, recurrent architectures, and released checkpoint populations. The 12 released checkpoints do not preserve training-seed or parent-run lineage, whereas the 60-checkpoint replication supplies five explicit seed lineages per configuration.

Visual or discrete control, stochastic transition kernels, and real-robot settings are outside the current evidence.

E Reproducibility Details

Experiments use Python 3.10.20, PyTorch 2.4.1 with CUDA 12.1, NumPy 1.24.4, pandas 2.2.2, SciPy 1.13.1, Gymnasium 0.29.1, and MuJoCo 2.3.3. Compute consists of four NVIDIA RTX 3090 GPUs with 24 GB memory.