REPLAY-CONDITIONED EXPLANATIONS: DISENTANGLING CURRENT, HISTORICAL, AND CONSTRAINT EFFECTS IN MEMORY-BASED REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Explaining a recorded decision of a memory-based reinforcement learning policy requires separating current observations, retained history, and action restrictions. We introduce replay-conditioned explanations, which bind one target decision to its policy and replay context, then evaluate these three interventions separately. The contribution of this paper is an executable query contract and an empirical account of the distinction between computational validity and reference-dependent diagnosis. Reconstruction checks and optional exact Shapley attribution make the explained computation explicit. Controlled mechanisms, reward-trained recurrent policies at 8 and 64 steps, and a native POPGym task support query correctness; public TimeSHAP agrees with matched exact event attributions. On the public task, all three fixed training seeds solve 51-decision episodes with a three-step memory dependency. Independent-reference tests expose a diagnostic limit: recovery falls from 100% with a clean paired reference to approximately 52% with the specified retrieved references. A calibration-frozen gate reduces clean modifications from about 48% to 4–5%, but recovery falls to 16–17%; greedy search matches exhaustive recovery with fewer calls. These results separate valid intervention computation from reliable fault diagnosis. They support explicit query and reference checks, while leaving broad benchmark generalization and new industrial deployment validation open.

1 INTRODUCTION

Memory-based reinforcement learning (RL) policies act on both current observations and retained history. A policy that stops near a wall may respond to visible geometry, retain an earlier cue, or face a constraint that removes a movement action. Explaining the recorded decision requires distinguishing these possibilities while keeping its action target fixed.

Feature and temporal attribution methods provide useful estimators (Lundberg & Lee, 2017; Bento et al., 2021; Leung et al., 2023), but their interpretation depends on the computation being explained. A current-only surrogate can match recorded actions through a proxy for history; editing history without reconstructing memory can hide its effect; changing legal actions can alter the output without changing sensory evidence. These are query ambiguities, not defects intrinsic to an estimator. Our public TimeSHAP experiment correctly separates current and earlier events under a matched query.

A second ambiguity concerns the quantity. Leave-one-out (LOO) removal measures a response conditional on the remaining factual inputs; Shapley averages marginal contributions across coalitions. Group-, entity-, and feature-level allocations may share probability units while using different player sets and references. They cannot be pooled merely because their units match.

We introduce *replay-conditioned explanations*: a fixed-decision contract binding the policy, schema, target action, legal set, and pre-step memory. It separates current replacement, history reconstruction, and constraint changes, while recording whether the declared query passed its checks. An optional exact-Shapley module attributes current-input effects at separately defined granularities.

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.

Neither numerical consistency nor a plausible-looking attribution certifies that a replacement reference is appropriate.

The contribution of this paper is an executable distinction between query validity and reference-dependent diagnostic utility. We test that distinction through controlled mechanisms, reward-trained recurrent policies, a public memory benchmark, and independent-reference diagnostics. The central question is when a correctly evaluated intervention supports diagnosis: replay checks can expose invalid computations, while an uninformative reference cannot identify a hidden clean state. Public TimeSHAP succeeds under matched queries; we claim neither a new attribution estimator nor a systematic defect in existing explainers. All training seeds are retained. Earlier team reports illustrate the industrial interpretation contract, without a new deployment experiment.

2 Related Work

SHAP allocates output differences over coalitions; Integrated Gradients integrates gradients along a reference path (Lundberg & Lee, 2017; Sundararajan et al., 2017). TimeSHAP explains sequence events, features, and cells with optional pruning (Bento et al., 2021). WinIT measures distributional effects over prediction windows using conditional replacement (Leung et al., 2023). These estimators address how to score contributions; our contract specifies which recorded policy computation is queried: fixed action and legality, factual or reconstructed memory, and declared replacements (Table 1). It can host temporal attribution and adds replay/reference checks, rather than a new Shapley estimator or a claim of superior attribution quality.

Table 1: Different methodological roles. Query binding names the supplied methodological role, not a capability ranking. All methods require a valid policy wrapper; explicit replay checks can accompany any estimator. Completeness is relative to each method's defined output/reference.

Method	Temporal scope	Additive accounting	Query binding
TimeSHAP	Events/cells	Shapley efficiency	User-defined wrapper
WinIT	Prediction window	Not a Shapley sum	User-defined wrapper
Integrated Gradients	Input trajectory	Path completeness	User-defined wrapper
This protocol	Current/history/mask	Within declared exact game	Explicit replay contract

For example, fixing target memory defines a current-input query; editing a past observation and replaying its suffix defines a historical query. Legality changes require a separate query even with the same estimator. The matched TimeSHAP experiment tests this compatibility (Section 5.1); WinIT has not been run.

RL explanations span recurrent-agent perturbations, causal accounts, interpretable policies, and future outcomes (Greydanus et al., 2018; Madumal et al., 2020; Wabartha & Pineau, 2024; Ruggeri et al., 2025). Saliency checks motivate computation validation (Adebayo et al., 2018; Yeh et al., 2019); masking independently changes policy outputs (Huang & Ontañón, 2022). Our conditional responses do not establish environmental causality or return gains. Appendix N covers further related work.

3 Preliminaries and Explanation Target

A recording is $D = \{(x_t, \ell_t, a_t, z_t)\}_{t=1}^{T}$, where x_t is the native observation, ℓ_t the legal-action set, a_t the recorded action, and z_t auxiliary inputs such as previous actions or rewards. A version bundle v binds the policy, encoder, schema, and semantic names. Reconstructed event logs select diagnostic contexts but need not be policy inputs. With state m_t immediately before x_t, a policy step is

$$(p_t, m_{t+1}) = F_\theta(x_t, \ell_t, m_t, z_t; v). \tag{1}$$

The state includes all recurrent/cache tensors and required reset and position metadata. For a fixed frame, head, action, and upstream-head conditioning $q = (t^*, k, a^*, c^*)$, define

$$f_q(x, \ell, m) = F_\theta^k(a^* \mid x, \ell, m, z_{t^*}, c^*; v), \qquad p^* = f_q(x_{t^*}, \ell_{t^*}, m_{t^*}). \tag{2}$$

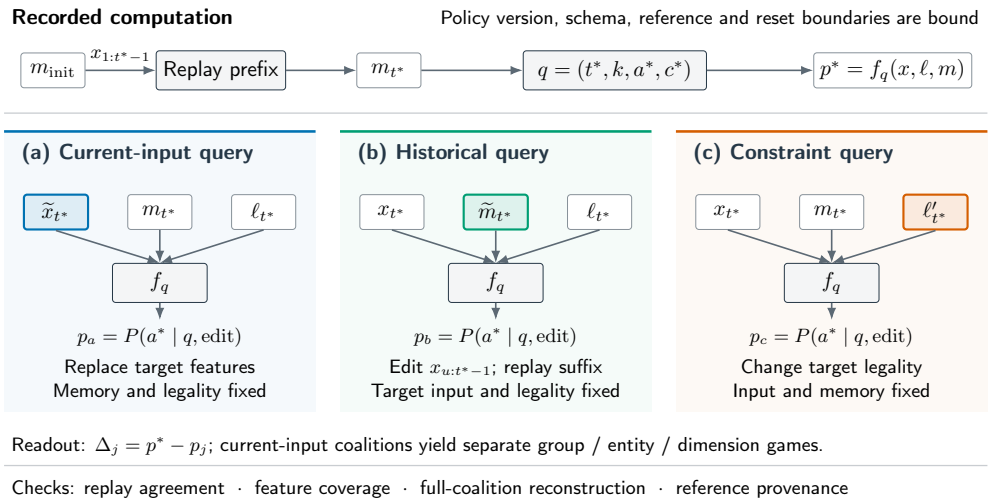

Figure 1: Replay-conditioned computation. Colored arguments are edited; white arguments remain factual. All queries share the recorded target and factual output. Historical edits require suffix replay. Current-input coalition games are separate from the three response queries.

The target action and c^* remain fixed under perturbation; recomputing upstream actions defines another query. An export that cannot fix this conditioning cannot certify the declared response. Each query also declares its scope and replacement $b \in \mathcal{B}$. Its output is a conditional probability difference. Numerical replay accuracy and reference plausibility are separate: even exact evaluation can use a physically inconsistent replacement.

4 Replay-Conditioned Explanations

Figure 1 separates the three queries; Algorithm 1 combines replay and query construction, with optional exact input attribution. Blue, green, and orange denote current, history, and constraint effects.

4.1 Three Fixed-Decision Queries

Replay the native record in order, retaining the required pre-step states. Every query receives an independent copy, including reset and cache-position metadata. For current group g, replacement $I_{g,b}$ holds history and legality fixed:

$$\Delta_{\text{cur}}(g, b; q) = p^* - f_q(I_{g,b}(x_{t^*}), \ell_{t^*}, m_{t^*}). \tag{3}$$

This LOO response is conditional on other factual inputs; a positive value means that the factual group supports the selected probability relative to b.

For a historical window $W = [u, w]$, $w < t^*$, edit the specified observations and replay the entire suffix:

$$m'_{t^*} = \mathcal{R}_\theta(m_u, I_{W,b}(x_{u:t^*-1}), \ell_{u:t^*-1}, z_{u:t^*-1}; v), \tag{4}$$

$$\Delta_{\text{hist}}(W, b; q) = p^* - f_q(x_{t^*}, \ell_{t^*}, m'_{t^*}). \tag{5}$$

The target observation, recorded auxiliary actions, and rewards stay factual. This is an open-loop memory response; the resulting state–action pair may be counterfactual. Directly editing the cache is a different query. Segment-freeze uses the observation preceding the interval; sparse timestep-freeze uses each selected step's predecessor.

Changing only target legality gives

$$\Delta_{\text{mask}}(\ell'; q) = p^* - f_q(x_{t^*}, \ell', m_{t^*}). \tag{6}$$

For binary post-logit masking and a nonempty legal set,

$$p_\ell(a) = \ell(a)e^{u_a} \Big/ \textstyle\sum_j \ell(j)e^{u_j}. \tag{7}$$

Relaxation changes normalization; legality-fed networks add another pathway. A relaxed-mask probability is a computational counterfactual, not automatically a learned preference (Huang & Ontañón, 2022). Historical mask edits require a separately declared query.

4.2 Optional Exact Attribution at Three Input Granularities

For finer current-input analysis, an optional module defines separate Shapley coalition games over observation groups, entities, or selected dimensions switched synchronously across entities. A player is one replaceable input unit; a coalition is the subset retained as factual. For layer L with players N_L, $x_L^{S,b}$ retains players in S and replaces other players by b; inputs outside the layer remain factual. With the same target, memory, and legality,

$$V_{L,b}(S) = f_q(x_L^{S,b}, \ell_{t^*}, m_{t^*}) - f_q(x_L^{\emptyset,b}, \ell_{t^*}, m_{t^*}), \tag{8}$$

$$\phi_i^{L,b} = \sum_{S \subseteq N_L \setminus \{i\}} \frac{|S|!(|N_L| - |S| - 1)!}{|N_L|!} \big[V_{L,b}(S \cup \{i\}) - V_{L,b}(S)\big]. \tag{9}$$

These are fixed-reference Shapley values with nonplayers held factual, not conditional-expectation Shapley values. Exact enumeration satisfies the within-game accounting check

$$\textstyle\sum_{i \in N_L} \phi_i^{L,b} = V_{L,b}(N_L). \tag{10}$$

Changing layers changes players, scope, and potentially the empty-coalition reference. Allocations are therefore not pooled across layers or added to Δ_{hist} and Δ_{mask}. The three-component restoration search in Section 5.3 does not use these allocations; its recovery gains do not establish this optional module's diagnostic value. LOO feature proportions are not fine-feature Shapley shares. The report's eight-group and four-entity games require 256 and 16 coalitions; selected-dimension games use 2^K, capped at $K = 10$ (Appendix G).

4.3 Validation and Reporting

Validate version/schema binding, covered inputs, reference provenance, and fixed conditioning before evaluation. Unknown uncovered inputs, including zero-valued fields, cause rejection. Check full-coalition tensor/output reconstruction and Equation 10; retain invalid and uncomputed results as such, never as zero effects. References record matching and future-information use; a within-recording mean can violate feature constraints. For joint replacements, $A(\{i,j\}) - A(\{i\}) - A(\{j\})$ measures nonadditivity when $A(S) = p^* - f_q$ under replacement set S.

Action agreement screens replay but cannot certify response accuracy. If factual and perturbed probability errors are bounded by ϵ_0, ϵ_1, the response error is at most $\epsilon_0 + \epsilon_1$ (Appendix H). Retrospective industrial reports provide no such bound. Every output records target, operator, layer/players, reference, versions, and validation status; Appendix D documents implementation limits.

5 Evaluation

We distinguish constructed policies, reward-trained agents, supervised diagnostic models, and retrospective reports; counts are not pooled. Appendices retain full protocols, seed outcomes, and failures.

5.1 Query Correctness and Implementation Validation

A three-step record has history h, current cue x, and memory $m = h$. Logits $(\kappa(\alpha x + \beta m + \gamma x m), 0)$ specify current, memory, additive, interaction, or constraint dependence. Three strengths, four cue pairs, and three legal sets give 180 configurations. All 540 responses match analytic probabilities within 2.22×10^{-16}. Reusing the unchanged cache after history edits fails on all 36 active unmasked cases (mean error 0.2727).

Algorithm 1 Replay-conditioned queries and exact input allocation

Require: Record D, versions v, fixed target $q = (t^*, k, a^*, c^*)$, query, scope, reference b
Validate bindings, input coverage, reference, and fixed head conditioning
Replay prefix; retain m_{t^*} and any required earlier state m_u
if validation or required replay checks fail **then**
    return invalid status and diagnostics
end if
$p^* \leftarrow f_q(x_{t^*}, \ell_{t^*}, \mathrm{copy}(m_{t^*}))$
if exact current-input allocation over N_L **then**
    if enumeration exceeds cap **then**
        return not-computed status and reason
    end if
    for $S \subseteq N_L$ **do**
        Build $x_L^{S,b}$, keeping nonplayers factual
        $P(S) \leftarrow f_q(x_L^{S,b}, \ell_{t^*}, \mathrm{copy}(m_{t^*}))$
    end for
    Check measured full-coalition reconstruction; compute $\phi^{L,b}$ by Eq. 9
    Check Eq. 10; **return** allocation and full provenance, retaining failures
end if
$(\bar{x}, \bar{\ell}, \bar{m}) \leftarrow (x_{t^*}, \ell_{t^*}, \mathrm{copy}(m_{t^*}))$
if current LOO query **then**
    $\bar{x} \leftarrow I_{g,b}(x_{t^*})$
else if historical query $W = [u, w]$ **then**
    $\bar{m} \leftarrow \mathrm{copy}(m_u)$
    for $j = u, \ldots, t^* - 1$ **do**
        $\tilde{x}_j \leftarrow I_{W,b}(x_j)$ if $j \in W$, else x_j
        $(_, \bar{m}) \leftarrow F_\theta(\tilde{x}_j, \ell_j, \bar{m}, z_j; v)$
    end for
else if constraint query **then**
    $\bar{\ell} \leftarrow \ell'$
end if
return $p^* - f_q(\bar{x}, \bar{\ell}, \bar{m})$ with query, reference, versions, and diagnostics

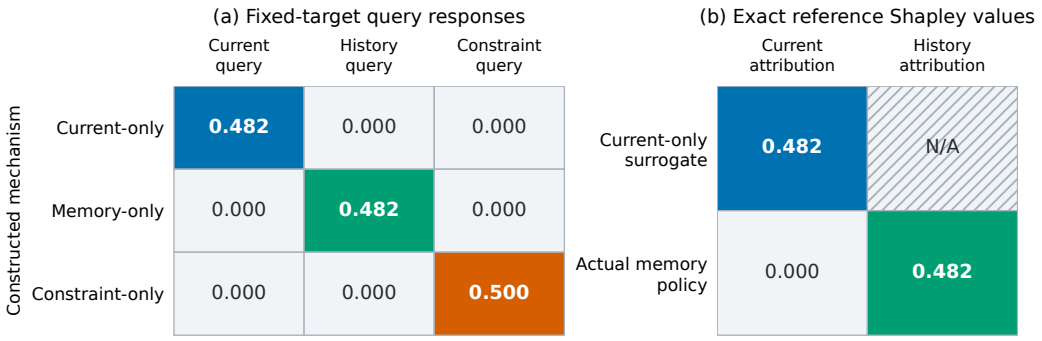

Figure 2: Separating query types and explained models. (a) Explicit responses for three isolated mechanisms ($h = x = 1$, $\kappa = 4$): 0.000 is a measured zero. (b) The current-only surrogate attributes a history proxy to current input; history is not an input to that surrogate (hatched N/A). Both panels use probability units but different estimands. This constructed example does not establish a defect in temporal attribution tools.

Figure 2a separates specified mechanisms, not classified samples. In panel b, a surrogate matches the memory policy at $x = h$ but attributes 0.482 to current input. The policy assigns zero to current input and 0.482 to history: matching recorded predictions need not explain the same computation.

LOO and exact Shapley measure different quantities. Their finite-grid mean signed deletion areas both equal 0.4121, versus 0.2591 for random ranking; reversed-cue evaluation gives 12 LOO wins, 156 ties, and 12 losses. Another 405 games show mechanism-dependent top-player disagreement, not estimator superiority (Table 17; Appendices J and L).

Matched public TimeSHAP. TimeSHAP 1.0.4 uses a fixed action/mask, zero cue references, factual phase/mode metadata, no pruning, and exhaustive event coalitions. Errors against independent exact event attribution are 1.11×10^{-16} over 180 constructed cases and 1.12×10^{-8} over 96 eight-step learned cases. It correctly separates current and earlier events; constraint effects lie outside this fixed-mask game. This supports compatibility, not superiority or intrinsic temporal confusion. No exhaustive 64-event comparison is claimed.

Learned implementation checks. On the GRUs below, fixed targets, factual nonplayers, and full-coalition reconstruction yield 576 games from 192 decisions and 3,072 coalition rows. Allocation error against permutation enumeration is at most 8.94×10^{-8}. Synthetic entities and coincident group/dimension games validate code paths, not industrial entities or ONNX replay (Appendix M). Table 2 summarizes agreement.

Table 2: Computation agreement, not explanation quality. Counts have row-specific units; complete query-level breakdowns are in Table 21.

Validation	Count	Maximum absolute error
Analytic queries	540	2.22×10^{-16}
64-step replay queries	576	5.96×10^{-8}
Learned input games	576	8.94×10^{-8}
TimeSHAP cases: constructed / learned	180 / 96	1.11×10^{-16} / 1.12×10^{-8}

5.2 Learned Recurrent Agents

Delayed-cue policies. A 24-unit GRU learns by REINFORCE to use the first or last sign, selected by a mode bit. Eight-step training uses 1,200 updates of 128 episodes; the 64-step curriculum adds 400 updates each at lengths 8, 16, and 32 before 1,200 at length 64. All seeds (11, 29, 47) are retained and tested on 4,096 new episodes each. First-cue accuracy falls from 74.6% to 50.7% for the weaker seed; two seeds reach 100% (Table 3). Different budgets confound a horizon-only interpretation.

Table 3: Overall / first-cue accuracy (%) for every fixed seed. Last-cue accuracy is 100% in all six runs.

Horizon	Seed 11	Seed 29	Seed 47
8 steps	100.0 / 100.0	87.3 / 74.6	100.0 / 100.0
64 steps	100.0 / 100.0	75.0 / 50.7	100.0 / 100.0

Each horizon supplies 288 target/mask contexts and 864 queries. At 64 steps, current/history replay discrepancies are at most 5.96×10^{-8}/zero. Mean absolute current/history responses are $0.176/0.396$ in first-cue mode and $0.495/9.19 \times 10^{-8}$ in last-cue mode (Figure 3). Task labels do not guarantee learned dependence; zero cue references are outside binary training support.

Native public memory task. A 32-unit GRU trains by reward-only policy gradient on native POP-Gym 1.0.7 `RepeatPreviousEasy` (Morad et al., 2023), with seeds 11, 29, 47 and 900 updates of 64 episodes. Each 51-decision episode predicts the observation three positions earlier: a three-step dependency, not 51-step retention. All seeds reach return 1.000 on 256 held-out episodes each; same-deck random returns are $-0.489, -0.486, -0.489$. These are newly trained, not official pre-trained policies.

Across 64 separate query episodes per seed, four targets and three categorical references give 2,304 exact four-event games. Current and suffix-replayed responses match full replay within 1.19×10^{-7}; lag three always ranks first. Selecting under one categorical shift and evaluating another yields absolute response 0.9994–0.9995, versus 0.2500–0.2501 for uniform event choice. Attribution costs 16 sequence evaluations. Native actions are all legal, so the constraint query is identity; synthetic masks are separate (Appendix P).

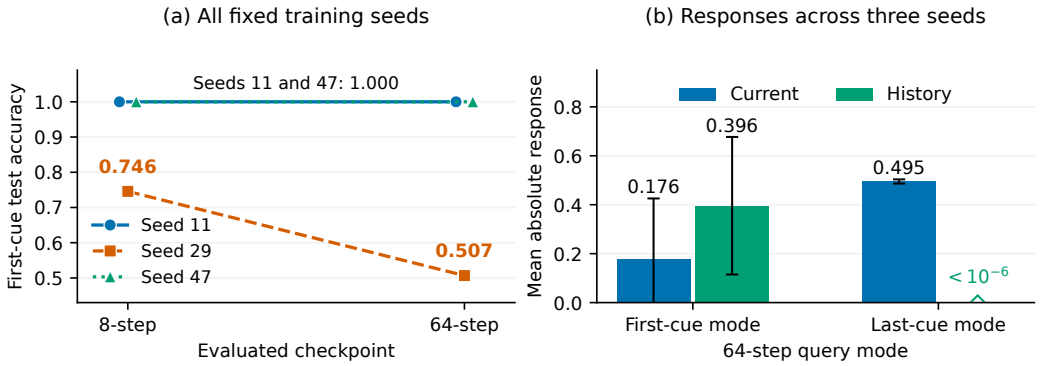

Figure 3: Reward-trained GRU results. (a) All first-cue accuracies; seeds 11 and 47 both equal 1.000. Markers are separated horizontally on the categorical axis; accuracy values are unshifted. (b) Mean absolute responses in unmasked 64-step contexts; error bars retain population standard deviations of three seed means. Different training budgets preclude a horizon-only comparison.

Held-out intervention and stale-cache ablation. Selection also transfers to predecessor-freeze: response 0.769–0.777, versus 0.192–0.194 for uniform choice and below 0.001 for the current event. No-op edits remain included (Figure 4a). These are sensitivity controls, not equal-cost estimators or return gains.

An injected adapter fault omits the last prefix observation. At identical targets, interventions, and six checked output evaluations per context, reconstruction checks reject all 768 faulty contexts and retain all 768 clean ones. Balanced-mixture coverage falls from 1 to 0.5 and accepted-query error from 0.5 to 0; risk is undefined in the fully rejected fault stratum. Both conditions compute validation outputs; gate-off ignores the checks. This isolates screening, not a natural package defect, speedup, or repair (Figure 4b).

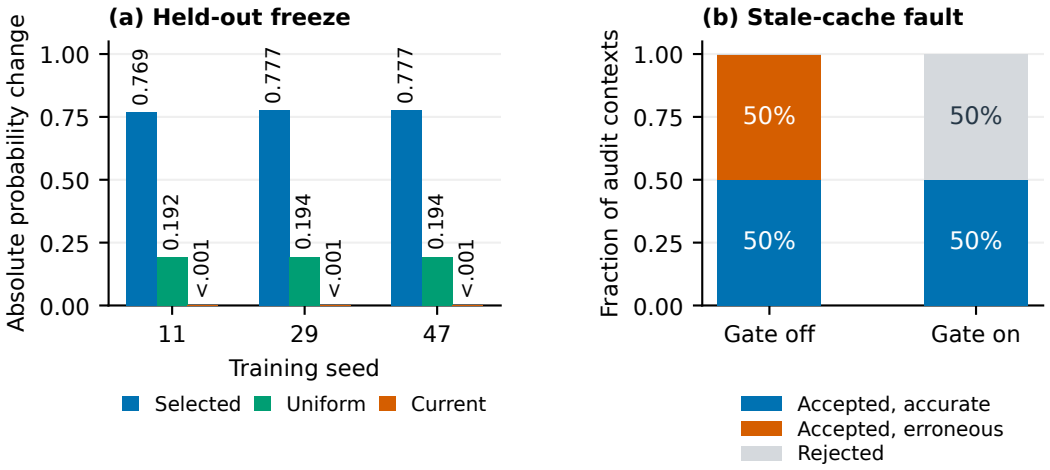

Figure 4: Public-task evidence. (a) Held-out predecessor-freeze sensitivity for every seed; uniform event choice averages all four candidate responses. These are sensitivity controls, not a matched-cost estimator comparison. (b) The same-query, same-computation adapter-fault ablation: rejected contexts are displayed separately from accepted accurate responses.

5.3 Reference-Dependent Diagnosis and Restoration

Known clean references. A separate supervised 16-unit, four-action GRU uses eight-step sequences and five seeds (all 100% test accuracy). Faults flip the current cue, substitute opposite-

cue memory, or mask the clean preferred action. On 6,000 single-fault and 2,000 clean contexts, full/history-only/current-only query F1 is 1.000/0.667/0.333; disabling the response/margin gate raises clean alerts from 0 to 1. These constructed, clean-pair ablations do not isolate every protocol check.

Searching eight replacement subsets for minimum size within total variation 0.02 of the clean reference recovers every changed action, versus 66.7% for independent replacement and $33.1 \pm 0.9\%$ for replacement-count-matched random (five-seed SD). Figure 5 exposes the mechanism: current+constraint and history+constraint faults hide the benefit of individual edits; joint replacement resolves them. Search averages 1.5 replaced components with no clean edits; replacing all three also recovers every fault but edits every clean context. Full replacement is feasible by construction. Random does not match policy calls, and input-granularity Shapley does not select repairs. These are ideal-reference single-decision results, not deployment or return gains (Appendix Q).

	C	H	K	C+H	**C+K**	**H+K**
Independent edits	1.000	1.000	1.000	1.000	**0.000**	**0.000**
Joint subset search	1.000	1.000	1.000	1.000	1.000	1.000
Matched random	0.350 ± 0.028	0.321 ± 0.027	0.338 ± 0.023	0.313 ± 0.016	0.324 ± 0.013	0.341 ± 0.023

Figure 5: Joint edits reveal masked effects under known clean references. C/H/K denote current/history/constraint faults. Cells show mean recovery over five seeds; nonzero population SDs appear below means. Recovery conditions on changed actions (200 episodes per fault set and seed). Independent edits fail on C+K/H+K. Random matches replacement count, not policy calls. Figure 17 gives full comparisons.

Independent reference retrieval. All six reward-trained checkpoints use 256 independent clean-bank, 128 calibration, and 128 test episodes each. Retrieval matches mode and distractors, excluding corrupted endpoint cues, test clean pairs, and labels. It therefore cannot identify the flipped relevant endpoint; this limitation concerns this retriever, not all reference construction. Clean-bank membership is assumed. Six fault sets plus clean cases, crossed with two edit families, give 10,752 contexts from 768 test episodes; legality faults are synthetic post-logit overlays.

Table 4: Independent-reference diagnosis, sign-flip faults. Within-episode averages, then equal-weight means of three seeds per horizon. Recovery uses only distribution-changing faults (TV > 0.05); clean edits use uncorrupted episodes. Both searches have an eight-query ceiling, not equal consumed cost. The gate threshold is frozen on clean calibration data. Full seed and episode-bootstrap results are in Appendix O.

Steps	Search / reference	Recovery	Clean edits	Edits†	Queries†
8	Exhaustive / retrieved	.516	.484	.614	4.61
8	Greedy / retrieved	.516	.484	.614	2.74
8	Gated exhaustive / retrieved	.164	.039	.212	2.15
8	Exhaustive / clean pair	1.000	.000	1.205	8.00
64	Exhaustive / retrieved	.518	.482	.614	4.63
64	Greedy / retrieved	.518	.482	.614	2.75
64	Gated exhaustive / retrieved	.170	.052	.218	2.19
64	Exhaustive / clean pair	1.000	.000	1.184	8.00

†Episode-averaged means on active faults; rejected cases retain zero edits and the screening query.

The clean-pair control recovers every active fault, but independent-reference recovery is approximately 52% (Table 4). Calibrating the clean-discrepancy gate reduces clean edits to 3.9–5.2%, while recovery falls to 16.4–17.0%. Greedy search matches recovery and replacement count with fewer distribution queries. Cached execution narrows the saving: exhaustive/greedy use 1.766/1.597 unique sequence evaluations at eight steps and 2.130/1.824 at 64, distinct from Table 4's query counts. These results establish a reference-dependent tradeoff, not deployment restoration.

Quality beyond the generating intervention. Rank the two cue endpoints by zero-replacement response, then evaluate only the selected endpoint by a sign flip, retaining the same fixed action. A randomized-ranking control has the same four-query budget (including the shared factual query). Across all six checkpoints, the increase in absolute held-out response is 0.242–0.508; each per-seed paired episode-bootstrap interval excludes zero. This supports sensitivity transfer across these two edit rules, not environment-level causality. Across four independently retrieved references, three-component attribution rankings have mean correlations 0.594–0.695, while sign agreement on components nonzero under either reference is only 0.424–0.473 for the sign-flip family, including clean cases. Stable replay therefore does not imply stable interpretation. Ties, zero effects, all seeds, and the separate zero-erasure family are retained in Appendix O.

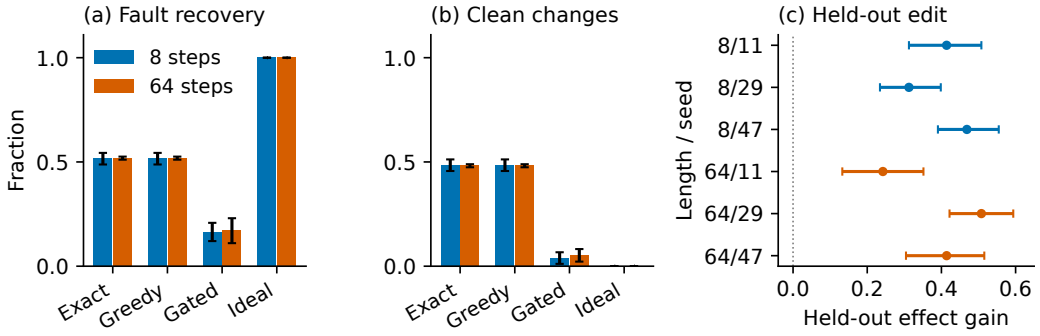

Figure 6: Independent-reference limits and held-out quality. (a,b) Recovery and clean edits; error bars are population SDs of three seed means. Exact and greedy coincide here; "gated" denotes gated exhaustive with the frozen calibration rule and "ideal" uses the clean pair. (c) Per-checkpoint effect gain under a second edit rule; intervals are paired 95% episode-bootstrap intervals, not uncertainty across tasks.

5.4 Retrospective Industrial Audit

Earlier team reports cover movement, button, and yaw heads. **Current input:** opening forward rays to 30 m changes few actions among 40 wall and 20 control frames in each of two episodes; at frame 193, yaw probability falls 0.110 without an argmax change (Table 14). **History:** target-switching surrogate traces motivate fixed-target replay queries, but those queries remain unexecuted (Figure 10). **Constraints:** unmasking changes winners in three selected cases; incomplete probability vectors preclude a fixed-action Δ_{mask} (Figure 13). Missing weights and trajectories prevent new deployment validation. Appendices F and B retain the report evidence and limits.

6 Discussion and Conclusion

The main finding is a separation: a declared query can be numerically correct while its reference supports unreliable diagnosis. Current, historical, and constraint interventions require different computations; replay checks expose deliberately invalid cached computations, and public TimeSHAP succeeds under matched semantics. Independent-reference experiments show that calibration reduces false alerts at a substantial recovery cost, and that exhaustive search is unnecessary on these tasks.

The public task adds native environment and reward evidence, but its memory dependency is only three steps. The 64-step task remains custom and includes a weak training seed. Reference pools assume known-clean membership; the specified retriever lacks information about corrupted endpoints. Synthetic masks and entity slots do not validate industrial constraints or multi-head attention caches. Three-granularity input attribution remains an optional, numerically verified module, with no demonstrated repair benefit. Public-task fault recovery, broader benchmarks, WinIT comparison, and environment-return improvements remain untested. The industrial evidence is retrospective. All seeds, failures, and source limitations are retained.

Reproducibility Statement

The accompanying experiment code includes the reward-trained GRU checkpoints, all seed results, public TimeSHAP comparisons, backend patch and regression checks, extended mechanism grids, audited diagnostic-study outputs, and the independent-reference and native POPGym studies. The latter studies retain disjoint RNG domains, frozen calibration rules, all three training seeds, checkpoints, policy-query accounting, and episode-clustered bootstrap summaries. The controlled-study code includes the complete finite-policy grid, exact coalition values, and analytic-response checks described in Appendix J. Appendix A specifies the reproduction record; Appendix K documents the retrospective source-selection rules and the revised report's methodological additions. Reproducing measurements for the industrial case study requires policy weights, native recordings, and output arrays that are outside the present package. The manuscript source reproduces the included tables and English figures.

Ethics Statement

The study concerns game-agent behavior. The accompanying artifacts contain constructed policies, newly trained task policies, and reported aggregates; industrial models, player recordings, and identifiable source traces are excluded. Interpretations concern the fixed policy computation specified in Section 3.

AI-Assistance Statement

In this work, we used generative AI tools for manuscript text polishing, and LaTeX syntax correction. We have not used generative AI tools for research idea formulation, mathematical derivation, experiment design, data collection, or result analysis. Model training, policy implementation, and empirical evaluation are not applicable AI-assisted tasks for this work. Additionally, we used generative AI tools for proofreading and rephrasing of supplementary appendix text. We have reviewed all AI-assisted work. All AI-modified text and pseudocode were manually inspected, revised, and validated by the authors to rule out unintended factual or logical errors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

References

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/294a8ed24b1ad22ec2e7efea049b8737-Abstract.html.

João Bento, Pedro Saleiro, André F. Cruz, Mário A. T. Figueiredo, and Pedro Bizarro. TimeSHAP: Explaining recurrent models through sequence perturbations. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2021. doi: 10.1145/3447548.3467166. URL https://arxiv.org/abs/2012.00073.

Thomas Bush, Stephen Chung, Usman Anwar, Adrià Garriga-Alonso, and David Krueger. Interpreting emergent planning in model-free reinforcement learning. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/cc4d9cfc45325e460b455a820d5f212c-Paper-Conference.pdf.

Jonathan Crabbé and Mihaela van der Schaar. Explaining time series predictions with dynamic masks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2166–2177, 2021. URL https://proceedings.mlr.press/v139/crabbe21a.html.

Samuel Greydanus, Anurag Koul, Jonathan Dodge, and Alan Fern. Visualizing and understanding Atari agents. In *Proceedings of the 35th International Conference on Machine Learn-*

ing, volume 80 of *Proceedings of Machine Learning Research*, pp. 1792–1801, 2018. URL https://proceedings.mlr.press/v80/greydanus18a.html.

Shengyi Huang and Santiago Ontañón. A closer look at invalid action masking in policy gradient algorithms. In *The International FLAIRS Conference Proceedings*, volume 35, 2022. doi: 10.32473/flairs.v35i.130584. URL https://arxiv.org/abs/2006.14171.

Kin Kwan Leung, Clayton Rooke, Jonathan Smith, Saba Zuberi, and Maksims Volkovs. Temporal dependencies in feature importance for time series predictions. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2107.14317.

Mingxuan Li, Kai-Zhan Lee, and Elias Bareinboim. Counterfactual shapley credit assignment. *Reinforcement Learning Journal*, 7, 2026. URL https://rlj.cs.umass.edu/2026/papers/Paper45.html.

Jiashuo Liu, Nabeel Seedat, Peng Cui, and Mihaela van der Schaar. Going beyond static: Understanding shifts with time-series attribution. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/50ea4cf434c84a50203d7145afc7b853-Paper-Conference.pdf.

Shicheng Liu and Minghui Zhu. UTILITY: Utilizing explainable reinforcement learning to improve reinforcement learning. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/548a482d4496ce109cddfbeae5defa7d-Paper-Conference.pdf.

Zichuan Liu, Yingying Zhang, Tianchun Wang, Zefan Wang, Dongsheng Luo, Mengnan Du, Min Wu, Yi Wang, Chunlin Chen, Lunting Fan, and Qingsong Wen. Explaining time series via contrastive and locally sparse perturbations. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/bd9b0b6ffab928b68c2b126d50f26381-Paper-Conference.pdf.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

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. doi: 10.1609/aaai.v34i03.5631.

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*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/70b8505ac79e3e131756f793cd80eb8d-Paper-Conference.pdf.

Steven Morad, Ryan Kortvelesy, Matteo Bettini, Stephan Liwicki, and Amanda Prorok. POPGym: Benchmarking partially observable reinforcement learning. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=chDrutUTs0K.

Ram Rachum, Yotam Amitai, Bálint Gyevnár, Reuth Mirsky, and Cameron Allen. Evaluating RL explainability methods by how much they help fix bugs in agents. In *Workshop on Explainable Artificial Intelligence at IJCAI-ECAI*, 2026. URL https://arxiv.org/abs/2608.17524.

Dilina Rajapakse, Juan C. Rosero, and Ivana Dusparic. TREX: Trajectory explanations for multi-objective reinforcement learning. *arXiv preprint arXiv:2603.21988*, 2026. URL https://arxiv.org/abs/2603.21988.

Rishav Rishav, Somjit Nath, Vincent Michalski, and Samira Ebrahimi Kahou. Behaviour discovery and attribution for explainable reinforcement learning. *arXiv preprint arXiv:2503.14973*, 2025. URL https://arxiv.org/abs/2503.14973.

Franco Ruggeri, Alessio Russo, Rafia Inam, and Karl Henrik Johansson. Explainable reinforcement learning via temporal policy decomposition. *arXiv preprint arXiv:2501.03902*, 2025. URL https://arxiv.org/abs/2501.03902.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328, 2017. URL https://proceedings.mlr.press/v70/sundararajan17a.html.

Ahmad Terra, Mohit Ahmed, Rafia Inam, Elena Fersman, and Martin Törngren. Textual explanations and their evaluations for reinforcement learning policy. *arXiv preprint arXiv:2601.02514*, 2026. URL https://arxiv.org/abs/2601.02514.

Sana Tonekaboni, Shalmali Joshi, Kieran Campbell, David K. Duvenaud, and Anna Goldenberg. What went wrong and when? instance-wise feature importance for time-series black-box models. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/08fa43588c2571ade19bc0fa5936e028-Abstract.html.

Maxime Wabartha and Joelle Pineau. Piecewise linear parametrization of policies: Towards interpretable deep reinforcement learning. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/1e75f7539cbde5de895fab238ff42519-Paper-Conference.pdf.

Zhuorui Ye, Stephanie Milani, Geoff Gordon, and Fei Fang. LICORICE: Label-efficient concept-based interpretable reinforcement learning. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/ff0e299b1cd800f006f3516f06177e2a-Paper-Conference.pdf.

Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David I. Inouye, and Pradeep K. Ravikumar. On the (in)fidelity and sensitivity of explanations. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/a7471fdc77b3435276507cc8f2dc2569-Abstract.html.

Borui Zhang, Wenzhao Zheng, Jie Zhou, and Jiwen Lu. Path choice matters for clear attributions in path methods. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/1dc2fe8d9ae956616f86bab3ce5edc59-Paper-Conference.pdf.

Xu Zheng, Farhad Shirani, Zhuomin Chen, Chaohao Lin, Wei Cheng, Wenbo Guo, and Dongsheng Luo. F-Fidelity: A robust framework for faithfulness evaluation of explainable AI. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/2231d1abdcc68d16c9861a58f3c25849-Paper-Conference.pdf.

A Minimum Reproduction Record

Keep units distinct: probabilities must concern the same head/action; surrogate SHAP, logits, and margins retain their own scales. Uncertainty reports must distinguish episode sampling, reference variation, and attribution sampling.

B Historical Behavioral and Surrogate Analyses

Historical circling, capture, rescue, and switching values are transcribed under Appendix K. Unknown SHAP links, held-out fidelity, episode membership, and uncertainty prevent interpreting surrogate units as policy effects.

Table 5: Required provenance for reproduction. This is a record specification; the present draft does not claim that every industrial asset has already been packaged in this form.

Record component	Required contents
Model and semantics	Policy hash, export version, encoder and capability-pack versions, input/output schema, action-head definitions
Recorded trajectory	Recording hash, episode and event identifiers, native/reconstructed provenance, timestamps, reset boundaries, auxiliary inputs
Replay configuration	Runtime and precision, sampling/temperature settings, cache initialization and truncation, available comparison outputs, errors and exclusions
Explanation target	Target frame, head, fixed action, scalar response, current group or historical window, multi-head conditioning
Reference	Pool and split, reference identifiers, matching rule, whether future information is used, semantic/temporal checks
Result	Original and perturbed outputs, state/checkpoint references, raw-logit availability, diagnostics, cost, failure status
Reporting	Evidence type, supported interpretation, uncertainty scope, unresolved validity checks, figure/table source

Table 6: Reported proximity-cue measurements across selected phases. The first three windows are nested. Endgame control/circling labels select the behavior by construction.

Phase	Cue mean	Nonzero (%)	Surrogate SHAP	Circling (%)
Opening 0–25	0.01	1.25	0.299	0
Opening 0–50	0.22	22.25	0.202	0
Opening 0–100	0.55	55.35	0.115	0
Midgame 200–600	1.00	99.60	0.012	1.18
Endgame control	0.85	85.25	0.067	0
Endgame circling	0.51	51.14	0.190	100

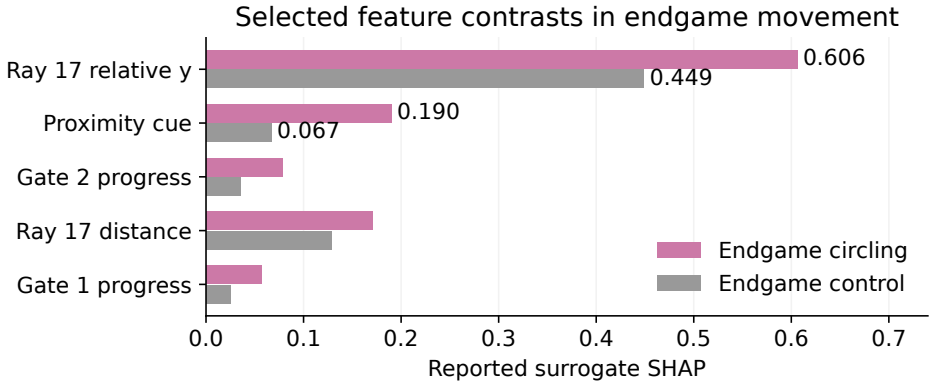

Figure 7: Selected movement-surrogate feature contrasts. The proximity cue changes from 0.067 to 0.190; ray-17 relative height changes from 0.449 to 0.606. The contrast motivates a phase-matched cue intervention.

B.1 Circling and Phase-Conditioned Features

Circling means a 20-frame displacement/path ratio below 0.35 or cumulative yaw above 270°: 231/562 endgame recordings (41.1%) among 644 recordings. Segment/all-phase totals are excluded.

Opening SHAP 0.299 coexists with zero circling. Nested windows are descriptive; constant elapsed time and feature-scale differences warrant schema checks.

B.2 Capture-Site Behavior and Conditional Associations

Adjacent tables assign counts 88/235 to near/far sites versus departure with/without nearby targets. Figure 8 retains near/far labels; event IDs are needed to reconcile them.

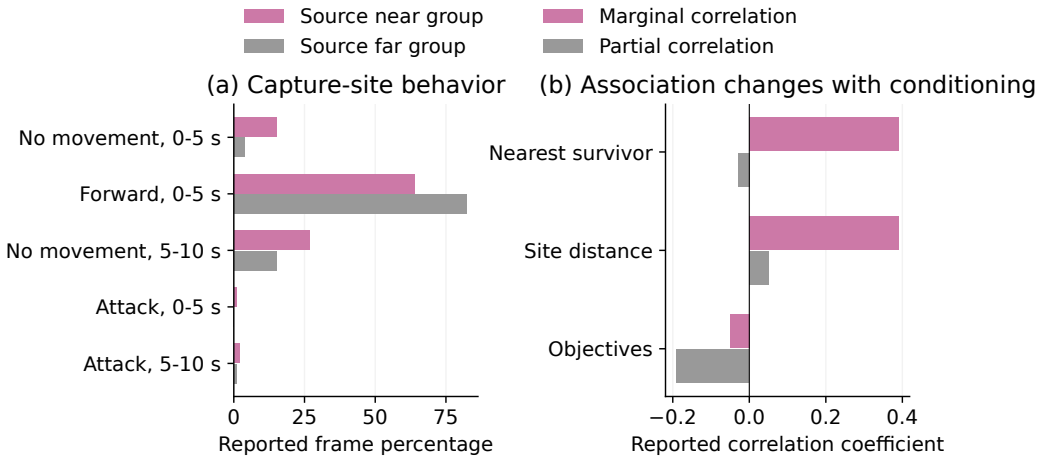

Figure 8: Historical capture-site summaries. Left: frame-level behavior percentages under the source group labels. Right: marginal and partial correlations with the reported departure outcome. The abbreviated right-axis labels denote nearest-survivor distance, survivor-to-site distance, and remaining objectives.

Table 7: Random-forest importance for the reported departure outcome. Training/evaluation split and uncertainty are unspecified.

Feature	Reported importance
Nearest survivor distance	0.322
Survivor-to-site distance	0.299
Elapsed time	0.178
Nearby survivor	0.066
Remaining objectives	0.061
Free survivors	0.026
Remaining survivors	0.019
Visible survivor	0.015
Downed survivor	0.013

Reported logistic odds ratios are 1.90, 0.53, 0.63, 1.56 for survivor-to-site distance, remaining objectives, nearby survivor, and downed survivor ($p < 0.001$ each). Scaling, coding, sample, and episode clustering are unknown; Figure 8 shows conditioning-sensitive correlations.

Nearby-survivor surrogate attack probabilities are 0.757/0.635 after capture/in controls; predicted rates are 82.5/68.7%. Comparison with observed-action groups requires matched membership and held-out fidelity.

B.3 Rescue Outcomes, Recorded Attacks, and Surrogate Contrasts

Rescue is a captured-to-free transition with health ≤ 0.7. Six data sources report 342 events: 243 within 200 game units, 99 farther away, tracked for ten seconds (60 frames). Recording counts and cohort overlap are unknown.

Of 243 near events, $(48 + 12)/243 = 24.7\%$ involve rescued-character health loss. Of 129 no-loss events, 83 lack attack actions (64.3%) and 46 contain them. Legal-attack frame rates are approximately 30/30/25% for neither/rescuer-only/rescued-only loss; marginal rates do not establish overlap. Raw-event attribution and conflicting aggregate rates remain unresolved (Appendix K).

Table 8: Mutually exclusive reported health-change outcomes in the 243 near-rescue events. Percentages are recomputed from the listed counts.

Outcome	Events	Percentage
Neither character hit	129	53.1
Only rescuer hit	54	22.2
Only rescued character hit	48	19.8
Both hit	12	4.9
Total	243	100.0

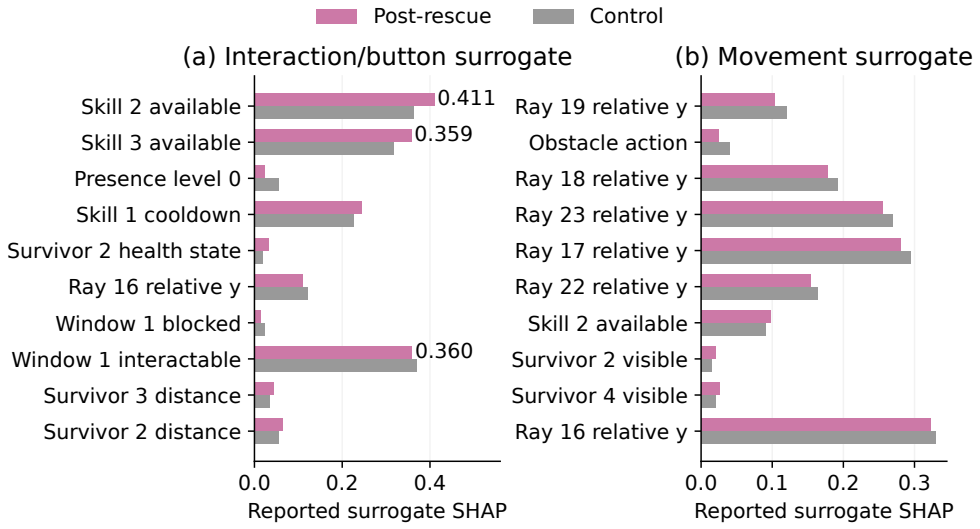

Figure 9: Historical post-rescue versus control surrogate SHAP values. The report samples 2000 frames per group, with episode membership unspecified. The two heads retain their respective output scales; numerically discussed bars are labeled.

Post-rescue interaction SHAP ranks skill/window features above health/capture history (Figure 9). Policy queries and exact long-horizon counts are unavailable; target preference remains untested.

B.4 A Target-Switching Case

In frames 400–520, visibility SHAP reaches $+0.66$ at 442 and zero near 460; the reported peak is 1.24 and distance values 0.05–0.07. Figure 10 specifies the unexecuted fixed-target test.

C Detailed Wall-Study Measurements

Episodes A and B use the same policy; C uses a different policy and supports observational comparisons only. These labels do not identify the unresolved M1/M2 series. Missing frames are excluded without a missing-at-random assumption. Wall/control rules use forward rays within 7 m or downward hits within 20 m / moving frames with forward rays beyond 15 m. In A, 348 wall frames form 73 intervals of at least two frames; 71 turn within five frames (97.3%), distinct from within-interval yaw. Front-ray height/distance differences are $+3.99/-3.99$; selected map pixels differ by 2.6–2.7. Partly ray-defined groups measure association, not attribution fidelity.

Frame 193's forward-ray yaw response is 0.220 despite low rank. Group and distance-only replacements differ. Missing-rate estimates are excluded; numeric successive-target traces remain unavailable (Appendices K, E).

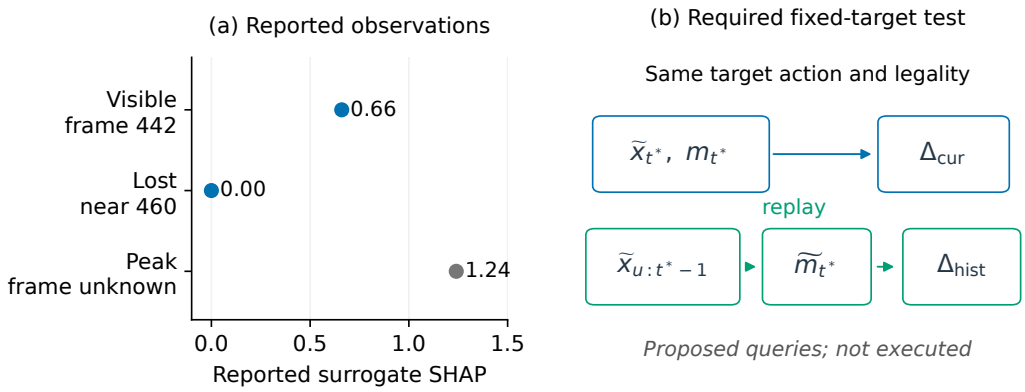

Figure 10: Reported evidence and unresolved mechanism in the historical switching case. Left: isolated visibility-surrogate values; the peak has no supplied frame index, and no trajectory is interpolated. Right: the fixed-target current/history queries needed to test the hypothesis, explicitly unexecuted. Values on the left are not policy-probability responses.

Table 9: Reported episode-level descriptive statistics. Wall and open-space percentages concern nonzero yaw outputs. Frame counts are not independent experimental replicates.

Quantity	A	B	C
Decoded frames	1779	1754	1622
Missing frames	382	430	445
Wall-adjacent frames	348	258	314
Wall-adjacent fraction (%)	19.6	14.7	19.4
Yaw turning: wall (%)	64.7	52.3	57.6
Yaw turning: control (%)	48.3	45.8	50.7
Difference (percentage points)	16.4	6.5	6.9
Rule override of yaw (%)	0.0	0.0	5.5

Table 10: Previously reported action changes after ray opening. Frames are dependent within each of two distinct episodes; collection independence is unverified.

Episode	Subset	Movement	Yaw
A	Wall-adjacent	0/40	1/40
A	Open-space	1/20	0/20
B	Wall-adjacent	1/40	0/40
B	Open-space	0/20	2/20

Table 11: Reported absolute feature-occlusion responses. The reference is the report's current-input analysis, not a historical intervention. Responses need not sum to one.

Frame	Movement head	Yaw head
193	Survivor bearing sine: 0.050; camera sine: 0.00007	Survivor bearing sine: 0.937; camera sine: 0.937
197	Survivor bearing sine: 0.778; camera sine: 0.048	Survivor bearing sine: 0.999; camera sine: 0.999
120	Board bearing sine: 0.00021; survivor distance: 0.00014	Skill charge: 0.986; survivor distance: 0.023

D Implementation Correspondence and Validation Scope

The anonymous snapshot contains `explainability_algorithm_bundle/model_explain_backend`, three coalition enumerators, a ten-player cap, and tag `shapley_g8_v1`. Original source, a separate patch, and controlled GRU checkpoints are retained; industrial weights, native trajectories, and new deployment outputs are unavailable. Table 12 maps computations without reassigning inherited figures.

Table 12: Implementation correspondence in the supplied source snapshots. Entries for the newly supplied v2 snapshot report code location and intended computation, not completed empirical validation.

Operation	Module / function	Scope
Version binding	`pack_registry.py`; layout checks	Schema, parser, feature names and model dimensions
Current-input exact allocation	`model_file_attribution.py`; `_run_shapley_layer`	v2 snapshot source for group-, survivor-, and selected-key-dimension coalition enumeration; no inherited figure is reassigned to this estimator
Fine-grained current-input localization	`model_file_attribution.py`; Top-20/decomposition paths	Single-feature LOO responses retained for localization; these values are not exact-Shapley shares
Historical response	`temporal_attribution.py`; segment/timestep attribution	Reconstruct state, hold one target observation fixed, preserve sparse replacement semantics
Replay screening	`temporal_orchestrator.py`; replay scan	Recorded model-action agreement; engineered global/window criteria
Reconstruction audit	`temporal_attribution.py`; reconstruction test	Tensor equality plus agreement of single, joint and residual effects
Reference comparison	`baseline_robustness.py`; ensemble	Direction, leave-one-out group and coalition consistency; uncalibrated classifications
Cross-recording references	`baseline_pool.py`	Model-identity eligibility, retrieval and overlap/diversity controls

Historical and current coverage. Historical wall occlusion excludes spatial maps; temporal code includes them and audits coverage. These backends use different references, and newer coverage does not extend the historical experiment.

Table 13: Implementation validation scope and current snapshot status. Test suites overlap and are not added together.

Artifact	Validation	Coverage limits
Original v2 native tests	98 passed; 8 protocol-dependent skips	Eight collectable files; the full nine-file suite fails collection because a platform module is absent
Original coalition arithmetic	Five deterministic checks pass	Weights, additive, redundant, and interaction games; no policy data
Original layer integration	Three semantic probes fail	Sign convention, fixed excluded groups, and full-coalition consistency
Constructed study	540 responses checked against analytic probabilities	Five specified mechanisms; no policy training

Missing platform/statistics dependencies block the ninth file; absent generated protocols cause the eight skips. Industrial replay additionally requires unavailable weights and recordings.

Arithmetic versus integration. The original arithmetic passes five deterministic checks at 10^{-12}, but stubbed probability adapters expose three integration failures:

1. **Sign.** Retained-coalition probabilities $(0.2, 0.4, 0.6, 0.8)$ imply $(+0.2, +0.4)$; using $p^* - p_S$ produces $(-0.2, -0.4)$, contradicting positive support and LOO.
2. **Fixed complement.** For $f(g_0, g_1) = 0.2 + 0.02g_0 + 0.01g_1$, factual $(10, 20)$, zero reference, and player g_0, holding $g_1 = 20$ gives endpoints $0.4, 0.6$. The backend zeros g_1, giving $0.2, 0.4$.
3. **Full coalition.** Overwriting the resulting full-coalition drop 0.2 with zero yields allocation -0.4 instead of $+0.2$; changing sign alone is insufficient.

The separate patch uses retained probabilities, fixes excluded inputs, validates full coalitions, and changes the estimator tag; failures and source hashes remain available.

Patched probes. Six patched tests check group support, pruning, survivor and dimension allocation, empty players, and rejection of invalid full coalitions. Group $(0.2, 0.4)$ and survivor $(0.15, 0.25)$ allocations agree within 4.77×10^{-8} (float32); five arithmetic checks also pass. These stubbed-adapter tests and learned-GRU checks (Appendix M) do not validate industrial ONNX execution. Native-suite counts refer to the original snapshot.

Other source-level limits. Above ten players, skipped entries serialize as zeros (uncomputed). The checkpoint tag incompletely binds provenance and is dropped from the final object. Missing legality defaults to all-legal; baseline argmax is not checked against the action output. Decomposition rescales absolute feature-LOO ratios by group magnitude; it is not signed feature/survivor Shapley. These limits require the checks in Section 4.

E Source-Display Provenance

Seven paired report figures are pixel-identical (Table 18); original displays and estimators remain in provenance assets. Only explicit values are redrawn.

M1/M2 identities and ray-to-metre conversion, the spatial-difference matrix, and successive-target trace arrays are unavailable. These displays remain source material, not fixed-target history evidence. Tables 9, 11, and 14 retain explicit values; Figure 13 marks untranscribed probabilities.

F Target-Bearing Intervention Details

Table 14 retains the source sign, $\delta p_{\text{src}} = p_{\text{changed}}(a^*) - p_{\text{base}}(a^*)$, opposite to the positive-support convention of Equation 3. L, R, and N denote left, right, and no turn; angles are degrees.

Table 14: Reported yaw actions and baseline-action probability changes after bearing rotations. The ray-opening column is a separate negative control; a preserved action is compatible with a nonzero probability response.

Frame	Baseline	$+90°$	$-90°$	$180°$	Rays open
193	L20	R40 (-0.937)	L40 (-0.902)	R40 (-0.937)	L20 (-0.110)
197	N	R40 (-0.999)	L20 (-0.684)	R40 (-0.999)	N (≈ 0)
120	L10	L40 (-0.984)	L10 (-0.146)	L40 (-0.953)	L10 (≈ 0)

Row-level recount. Rotations change $3 + 3 + 2 = 8$ of nine actions, contrary to source prose (Appendix K); frame 193 retains its ray-opening argmax despite a probability response.

Population and context. The three frames share one episode; wall studies separately use 60 frames per episode in two episodes. Joint sine/cosine edits preserve the unit circle but may violate spatial/dynamic relations: these are local policy responses.

G Input Allocation and Replay Cost

Algorithm 1 separates input-coalition allocation from direct current/history/constraint responses. Table 15 gives enumeration budgets.

Table 15: Current-input estimators specified by the revised report. Coalition counts follow directly from enumeration; they are not measured runtime or new deployment sample counts.

Player set / output	Computation and interpretation	Coalition budget
Observation groups	Exact allocation among eight groups	$2^8 = 256$
Survivor instances	Exact allocation among four instances	$2^4 = 16$
Selected dimensions	Exact allocation among K synchronous cross-instance switches	2^K, $K \leq 10$
Feature Top-20	Single-feature LOO response; localization	—
Feature decomposition	LOO-derived ratios anchored to a group Shapley magnitude; diagnostic only	—

Redundant inputs. As a simple illustration, let a fixed-action probability be $f(c_1, c_2) = 0.5 + 0.4 \max(c_1, c_2)$ for binary cues, with factual input $(1, 1)$ and zero reference. Each LOO response is zero, but exact Shapley assigns 0.2 to each cue, summing to the joint drop 0.4. Small single-feature sensitivity can therefore coexist with an additive allocation to redundant inputs.

Replay cost and state isolation. A history edit beginning at u costs a suffix of $t^* - u$ steps per reference/query. Pre-step checkpoints save prefix replay only with recording-bound hashes, initialization, and positions. Each branch copies mutable states and metadata; batching must isolate interventions. Table 23 reports measured costs for differing query budgets.

H Response Error and Identification Limits

The following elementary bound and counterexamples delimit the queries in Section 4.

A conditional response-error bound. Let $f_0, f_1 \in [0, 1]$ be the reference implementation's probabilities for the same target action before and after one specified intervention. Let $\hat{f}_0, \hat{f}_1$ be the corresponding replayed values. If $|\hat{f}_i - f_i| \leq \epsilon_i$ for both $i = 0, 1$, then

$$|(\hat{f}_0 - \hat{f}_1) - (f_0 - f_1)| \leq \epsilon_0 + \epsilon_1. \tag{11}$$

The triangle inequality gives the bound, also for other scores in their own units. Factual replay certifies only ϵ_0; intervened error ϵ_1 needs separate validation.

Replay screening thresholds. The available orchestrator uses recorded-action agreement of 0.99 globally, 0.95 in the target window, and exact agreement at the target. These are engineering defaults, not probability-error guarantees.

Action agreement does not certify response accuracy. Two implementations assign first-action probabilities $(0.9, 0.6)$ and $(0.6, 0.9)$ before/after intervention. Both always choose that action, but responses are $+0.3$ and -0.3, an error of 0.6. Thus an error bound cannot vanish with argmax disagreement alone.

Constraint-induced non-identification. With only action 1 legal, logits $(0, 10)$ and $(10, 0)$ both yield executed distribution $(1, 0)$ under Equation 7. The original executed distribution cannot distinguish these opposite unmasked preferences. A constraint query can reveal a different distribution, but identifying its meaning still depends on whether the mask acts solely after logits or also enters the network.

Current and historical queries are not interchangeable. For $f(x_t, m_t) = \sigma(m_t)$ and $m_t = x_{t-1}$, editing x_t has zero response; editing x_{t-1} and rebuilding memory can change it. Conversely, $f = \sigma(x_t)$ responds only to the current edit. Neither query alone establishes environmental causality.

I Remaining Comparative Evaluation

Planned, not executed. Multi-benchmark/long-horizon comparisons and industrial reruns remain open; the latter require weights and recordings.

Compare random/recency rankings, statistical screening, surrogate SHAP, current/window occlusion, TimeSHAP, aligned WinIT, and gradient-accessible sequence Integrated Gradients under identical targets, references, perturbations, and policy-call budgets. Apply the same validation checks to every estimator. Split by episode/generation session; freeze references and thresholds on calibration data, and report exclusions, versions, episodes, seeds, calls, replay steps, runtime, and peak memory. A pilot determines precision/sample size.

Test independently verified current/history/redundant/irrelevant dependencies. For fixed score f, use held-out deletion $D_k = f(x) - f(I_{S_k,b}(x))$, normalized-budget area, random rankings, and a second edit family. Report reference sign agreement, Spearman correlation, top-k overlap, and sparsity $k/|\mathcal{G}|$ at a fixed effect criterion. Keep surrogate prediction agreement separate from attribution fidelity and policy sensitivity separate from environment benefit.

Ablate version binding, memory propagation, context matching, and reference auditing; label injected faults. Coverage is accepted fraction, and selective risk requires independent correctness labels. Compare matched coverage with episode-clustered paired uncertainty; distinguish training seeds from attribution samples. Without correctness labels, report stability only.

J Constructed-Policy Study: Complete Protocol

The finite grid uses observations $(0, h, x)$ at steps $(0, 1, 2)$ and zero-initialized cumulative memory, so target memory is h. Logits $(\kappa(\alpha x + \beta m + \gamma x m), 0)$ precede a binary mask. One head is replicated across the adapter's three interface heads; no policy is trained.

Table 16: Specified mechanism families. Each contributes 36 configurations; the constraint-only family has equal pre-mask logits, so output probabilities are determined entirely by the legal set.

Family	α	β	γ
Current-only	1	0	0
Memory-only	0	1	0
Additive	1	0.5	0
Interaction	0	0	1
Constraint-only	0	0	0

Cross $\kappa \in \{1, 2, 4\}$, $h, x \in \{-1, 1\}$, and masks $\{0, 1\}, \{0\}, \{1\}$. Fix factual argmax (tie: action 0). Each of 180 configurations has three queries: zero current cue; replace step 1 by its zero predecessor and replay; allow both actions. Analytic maximum error is 2.22×10^{-16}. Unchanged-cache history edits average error 0.05454 overall and 0.2727 across 36 active unmasked cases (all fail).

Exact reference Shapley comparator. Players are current input, history, and constraint. For $S \subseteq N$, $|N| = 3$, the game value $v(S)$ retains original values for players in S, replaces other inputs by zero, and relaxes the constraint if its player is absent. It always evaluates the same target action. We enumerate all eight coalition values and compute

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - 1 - |S|)!}{|N|!} \big(v(S \cup \{i\}) - v(S)\big), \qquad |N| = 3.$$

An independent average over the six player permutations verifies this calculation and the efficiency identity to 10^{-14}. A direct response is instead $v(N) - v(N \setminus \{i\})$. Their different nonzero supports reflect different estimands, including interaction allocation.

Metrics and interpretation. Rank absolute scores, ties current/history/constraint. Evaluate with held-out replacements $x, h \mapsto -x, -h$ and relaxed constraints. Let D_k be factual minus perturbed target probability after the first k replacements; integrate the signed curve over $k/3$ by trapezoids. Averaging all six orders gives the exact random baseline.

Table 17: Original finite-grid results. Top-1 agreement compares two references over 168 active cases. Support is the fraction of three entries exceeding 10^{-12}; it is not a quality score.

Ranking	Mean signed area	Top-1 agreement	Mean support
Absolute direct response	0.4121	1.000	0.3556
Absolute exact Shapley	0.4121	1.000	0.6222
Uniform random order	0.2591	—	—

One reference change defines top-1 agreement; support is sparsity without matched deletion quality. Equal areas comprise 12 direct wins, 156 ties, 12 Shapley wins (10^{-12}), with no sampling inference. Direct queries use 13 policy-step calls versus eight Shapley coalition evaluations per reference; differing workloads imply no runtime advantage.

Proxy example and scope. A current-only logistic surrogate fits memory-policy logits at $x = h = \pm 1$ for all strengths. At $\kappa = 4$, $x = -h$ gives maximum error 0.9640. Six calibration states have positive surrogate current Shapley but zero policy current response; policy Shapley assigns history. This is surrogate omitted-variable error. Learned/public TimeSHAP tests are in Appendix L; WinIT and industrial reruns remain unexecuted.

K Report Provenance, Selection Rules, and Units

Both reports originate from our team's prior work: one supplies behavioral/surrogate analyses, the other selected native-policy interventions. We reanalyze these records; the separately reproducible controlled studies are new.

K.1 Summary-report versions and estimator provenance

All seven corresponding v1/v2 PNGs have identical hashes and zero decoded-RGB error. Version 2 adds exact group-, survivor-, and selected-dimension Shapley documentation, but the unchanged figures retain their original LOO, statistical, or counterfactual estimators (Table 18).

Table 18: Differences between the report versions for the industrial case study and how they are used here. Figure identity is based on pairwise PNG hashes and decoded-RGB comparison.

Item	v1	v2	Manuscript treatment
Seven report figures	Statistical, LOO, response-curve, and counterfactual displays	All seven pixel-identical to v1	Preserve original estimators; no new empirical result
Current-input allocation	Group/feature masking and probability drops	Exact group, survivor-instance, and selected-dimension coalitions	Revised capability, separate from inherited outputs
Fine-grained features	Single-feature LOO localization	Top-20 and decomposition remain LOO	Keep separate from exact-Shapley shares
Across-time display	Figure 4: current responses at successive targets	History replay is separate; Figure 4 unchanged	Successive-target trace, not fixed-target history

K.2 Report populations

Table 19 distinguishes screening populations from direct interventions in two episodes; frames within an episode are not independent replicates.

Table 19: Source-report populations and their role in the retrospective analysis.

Analysis	Source scope	Role
Historical circling	644 recordings; 562 with an endgame phase; 231 flagged	Generate phase-specific hypotheses
Capture and rescue	88/235 source-labeled capture groups; 342 rescue events	Compare behavior, conditions, and outcomes
Target switching	One illustrated episode	Identify a visibility hypothesis
Wall-ray intervention	40 wall and 20 control frames in each of two episodes	Test a current-input replacement
Constraint and bearing cases	Three displayed unmasking cases; three bearing frames	Inspect action restrictions and head-specific responses

Explicit table endpoints and category counts take precedence over conflicting prose; duplicated rows are counted once. This transcription rule does not resolve the underlying discrepancies without raw logs.

Table 20: Source discrepancies and the data retained for this analysis.

Source item	Discrepancy	Selection decision
Circling segments	320 in prose; 322 endgame and 569 overall in a screenshot	Exclude segment totals. Retain explicit recording counts 231/562; keep the all-phase count separate.
Episode C total	Prose 2065; table components 1622 and 445	Retain component counts; exclude prose total and missing-rate estimates.
Rescue attack rate	2.1% versus 0.9 frames per 60, implying 1.5%	Exclude both aggregate rate summaries; retain 83 and 46 of 129 events.
Bearing argmax count	Prose 7/9; displayed actions imply 8/9	Recount displayed actions as 3+3+2 changes; use row-level results for the case analysis.
Capture group labels	88/235 mean different groups in adjacent tables	Preserve source labels; exclude pooled contrasts and reconciled cohort claims.
Rounded SHAP differences	Some differences vary by 0.001 from endpoint subtraction	Use reported endpoints for plots and tables.
M1/M2 series	Episode and model mapping unavailable	Retain labels only in the inherited figure; exclude series-specific numerical conclusions.

Units. Wall rules use metres: horizontal/downward thresholds 7/20 m, controls beyond 15 m, and ray replacement at 30 m. Rescue uses 200 game-coordinate units and ten seconds (60 frames in that source). Distance-response curves use normalized rays, including a region below 0.035. Conversions are unavailable; thresholds remain specific to each analysis.

L Revision Experiments and Complete Measurement Scope

Controlled studies are separate from industrial records and the withdrawn proxy experiment; all seeds and failures are retained.

L.1 Reward-Trained Delayed-Cue Task

Eight-step observations contain a random sign, initial/final flags, and mode. Mode 0/1 rewards the first/last sign (positive/negative $\rightarrow$ action 0/1); intermediate signs distract. Only the terminal action earns ± 1, with no supervised loss or navigation.

A 24-unit GRU with a two-logit head trains by REINFORCE: batch-mean reward baseline, entropy coefficient 0.005, Adam rate 0.004, gradient cap 1, batch 128, and 1,200 updates for seeds 11, 29, 47. Evaluation uses single-thread CPU, deterministic PyTorch 2.12.0, and float32. Test/explanation seeds are training seed plus 10,000/20,000. Curves and all checkpoints are retained.

Table 21: Numerical validation against independent references. Rows have different units and are not pooled: queries for replay checks, games for allocations, and cases for TimeSHAP. Errors measure computation agreement, not explanation quality. The TimeSHAP masks are fixed outside its event game.

Checked computation	Count	Maximum absolute error
Analytic response queries	540	2.22×10^{-16}
64-step current queries vs. full replay	288	5.96×10^{-8}
64-step historical queries vs. full replay	288	0
Learned input games vs. permutation oracle	576	8.94×10^{-8}
TimeSHAP vs. exact event game: constructed	180	1.11×10^{-16}
TimeSHAP vs. exact event game: learned	96	1.12×10^{-8}

Table 22: All training seeds, evaluated on 4,096 newly generated episodes each. Counts retain label balance and the weaker policy's prediction imbalance. These are seed-specific outcomes, not independent estimates from pooled frames.

Seed	Overall (%)	First cue (%)	Last cue (%)	True action counts	Predicted counts
11	100.0	100.0	100.0	2034/2062	2034/2062
29	87.3	74.6	100.0	2031/2065	2553/1543
47	100.0	100.0	100.0	2052/2044	2052/2044

Two modes $\times$ two initial signs $\times$ two final signs $\times$ four distractor realizations give 32 sequences/seed; three masks give 288 contexts/864 queries. Target is factual masked argmax. Current/history edits zero last/first cues, keeping phase/mode. Cached/full replay agrees exactly. Singleton masks hide dependence revealed by relaxation; zero cues leave training support.

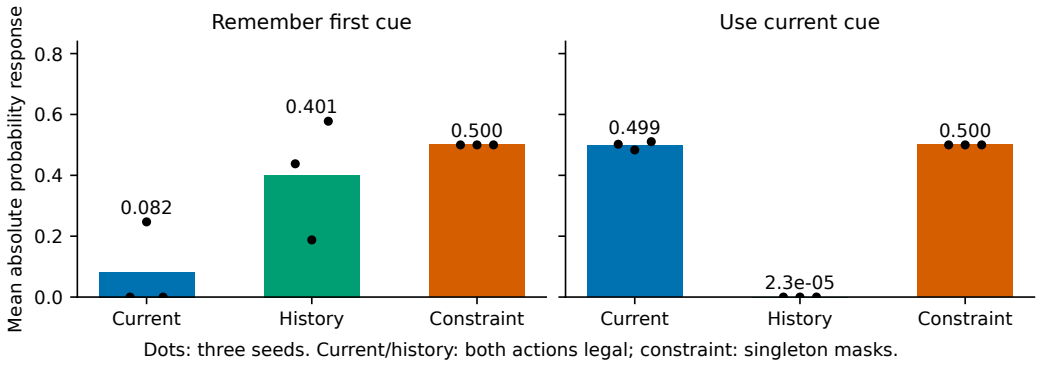

Figure 11: New reward-trained GRU experiments. Bars average absolute responses; black dots retain all three seed means. Current/history use both-legal contexts; constraints use singleton-mask contexts, so their heights describe different query sets.

Edit sensitivity. Segment $[2, 4]$/sparse $\{2, 4\}$ predecessor freezes differ by $1.95 \times 10^{-5}/1.45 \times 10^{-5}$ in unmasked first/final-cue modes; zeroing $[0, 2]$ versus $\{0, 2\}$ differs by $0.16065/7.73 \times 10^{-6}$. These are different edits.

Cost measurement. Table 23 assumes factual outputs/prefixes available. Direct calls process 1/8 steps; TimeSHAP batches 256 sequences in three calls. Direct timings: 20 warm-ups, 200 repeats/seed, median of seed medians. TimeSHAP includes construction over 96 cases. Fresh-process checkpoint-11 RSS includes interpreter/libraries/model. Runs use one CPU thread, ARM64 macOS; budgets differ.

Table 23: Measured CPU cost. Calls/steps counts model forward calls and total recurrent steps for the perturbed computation; factual prefix cost is excluded. RSS includes a fresh process's libraries and model. Workloads have different estimands and budgets.

Query	Calls / steps	Median ms	Peak RSS (MiB)
Current (cached)	1 / 1	0.032	264.8
History (replayed)	1 / 8	0.073	263.3
Constraint (target)	1 / 1	0.026	264.1
TimeSHAP events	3 / 2048	2.142	478.5

L.2 Public TimeSHAP Execution

Public TimeSHAP 1.0.4 is unchanged; a SHAP 0.51.0 compatibility alias maps `Kernel` to `KernelExplainer`. Settings are `pruning_idx=0`, seed 2027, `l1_reg=0`, and `nsamples=` 2^T. Wheel hashes, vectors, independent oracles, and logs are retained.

Constructed event games fix action/mask, with a dummy initial zero event and no constraint player. Learned games expose eight cues, reconstructing phase/mode. An independent enumeration/permutation oracle shares only the policy.

TimeSHAP shows no intrinsic current/history confusion; feature-level aggregation is not a temporal-error test. WinIT remains unexecuted.

L.3 Interaction and Proxy Extensions

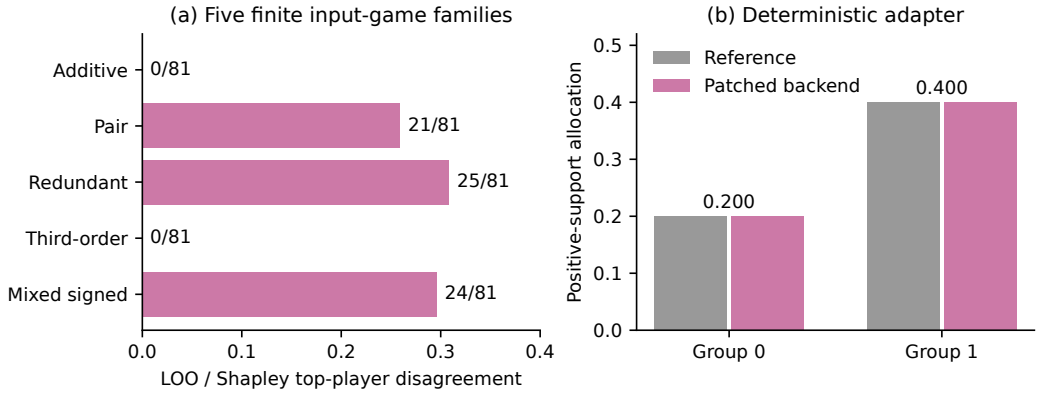

Figure 12: Executed allocation checks. Left: top-player disagreement across five input-game families. Right: positive group allocations from the patched backend against a deterministic adapter's analytic reference. The right panel is not a deployment result.

The new input games have three current-input players $x_i \in \{0.5, 1, 2\}$ and strength $\kappa \in \{0.5, 1, 2\}$, with probability $\sigma(\kappa g(x))$. A missing player is zero. The five functions are

$$g_{\text{add}} = 0.8x_0 + 0.5x_1 + 0.2x_2, \qquad g_{\text{pair}} = 0.8x_0x_1 + 0.3x_2,$$
$$g_{\text{red}} = \max(x_0, x_1) + 0.3x_2, \qquad g_{\text{third}} = x_0x_1x_2 + 0.2x_0,$$
$$g_{\text{mixed}} = x_0 + 0.5x_1 + 0.2x_2 - 1.5x_0x_1 + 2x_0x_1x_2.$$

Each family has 81 cases (405 total); rank absolute scores rounded to 12 decimals, ties by player index. Efficiency residual is 1.11×10^{-16}. Full-order disagreements are 0/21/59/0/26 of 81; Figure 12 reports top-player differences, not explanation superiority.

Current-only logistic surrogates fit two $x = h$ states per mechanism/strength. Across four (x, h) states and three strengths, surrogate/policy current-response discrepancies average 0, 0.3646, 0.1801, 0.3646, 0. This compares models, not Shapley implementations.

L.4 Transcribed Constraint Cases

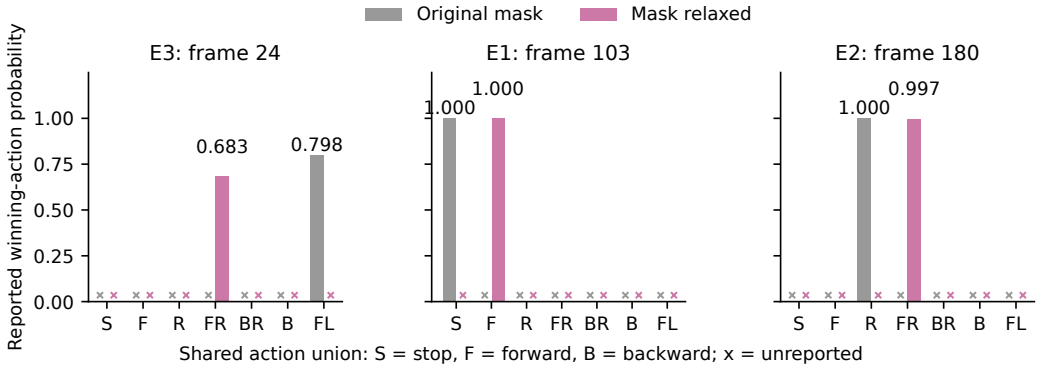

Figure 13: Prior-report winning-action probabilities under two constraint conditions. All panels share the same union of displayed action categories and ordering. Crosses mark probabilities not transcribed from explicit source numbers; they are not zero-probability bars or newly inferred values.

Displayed actions are not a complete legal set; missing probabilities are not inferred from complements.

M Longer-Horizon, Learned-Backend, and Supplied Diagnostic Studies

M.1 64-Step Reward-Trained Extension

The eight-step setup (Appendix L) trains fresh initializations for 400 updates each at lengths 8/16/32, then 1,200 at 64. This doubled budget confounds horizon comparisons; all seeds remain. For seeds 11/29/47, true counts are 2050/2046, 2033/2063, 2024/2072; predicted counts are 2050/2046, 2019/2077, 2024/2072. Seed 29 fails first-cue retention despite balanced predictions.

The unchanged query design gives 288 contexts (32 sequences per seed, three masks, three responses). Maximum current/full-replay discrepancy is 5.96×10^{-8}; historical reconstruction agrees exactly. Mean absolute current/history responses are 0.17628/0.39583 in first-cue mode and $0.49501/9.19 \times 10^{-8}$ in last-cue mode. For unmasked cases, segment/sparse predecessor-freeze differences over steps 2–4 are $8.69 \times 10^{-8}/1.74 \times 10^{-8}$ by mode; distinct zero-reference edits of $[0, 2]$ versus $\{0, 2\}$ give $0.22917/4.22 \times 10^{-8}$.

M.2 All Three Allocation Paths on Learned Policies

Patched constructors stub unrelated imports only. Six checkpoints $\times$ 32 new sequences (seed $+30{,}000$) give 192 decisions; factual argmax is fixed, both actions legal. Each coalition queries the trained GRU at the recorded prefix.

Channels form one cue group and three singleton context slots (initial/final/mode). Group allocation uses two players; entity allocation switches slots with cue fixed; dimension allocation switches context jointly and cue separately. Budgets 4/8/4 give 576 games/3,072 rows. Group/dimension games coincide; slots are synthetic, and the repeated three-head interface is counted once.

An independent full-sequence/permutation oracle gives maximum coalition/allocation/efficiency errors $1.79 \times 10^{-7}/8.94 \times 10^{-8}/5.42 \times 10^{-20}$. Full coalitions are checked; hashes and values retained. Zeroed phase/mode may leave support: agreement does not validate reference utility or industrial ONNX semantics.

M.3 Source Audit and Diagnostic Training

Report v3 pp. 17–22 consolidates existing controlled records (Table 24). Seventeen manifest differences are line endings; fresh hashes bind bytes. All 6,010 metrics agree within 3.33×10^{-16}. Seed-7 reruns match categorical outputs across 1,600 localization/8,400 restoration rows; numerical errors are $2.98 \times 10^{-7}/2.12 \times 10^{-6}$. Other seeds are row-audited. Studies use different test sizes/RNG consumption and are not pooled.

The supervised model is a 16-unit GRU/four-logit classifier: label $2b_h + b_c$ uses initial/final bits in channels 0/1; channels 2/3 are Gaussian distractors (SD 0.08). Train on 2,400 eight-step examples for 120 full-batch epochs, masked cross entropy, Adam 0.02. Masks retain target plus another action; tests start all-legal. Seeds 7/11/13/17/19 all achieve 1.000; abort threshold is 0.95, with no evidenced additional attempts. These are not reward-trained policies.

The excluded two-step study reduces its mixed label to current-only and maps one event to current/constraint labels, confounding ties. Its TimeSHAP-/WinIT-style variants are not public packages; Section 5.1 supplies public TimeSHAP results.

Table 24: Coverage of the supplementary experiment report (third revision, pages 17–22). Counts refer to the same supplied artifacts used above, not additional replications.

Layer	Reported scope	Treatment in this manuscript
E1	180 constructed contexts; 540 queries	Analytic verification in Section 5
E2	Two-step supervised semantic task; 8,000 examples	Excluded as comparative evidence for the label/mapping defects described here
E3	6,000 faulty + 2,000 clean contexts; 750 invalid queries	Section 5.3; Figure 14
E4	63,000 repeated-scenario contexts	Figures 15a and 16a
E5	42,000 repeated-scenario contexts	Section 5.3; Figure 17

M.4 Faults, References, and Scoring

Clean pairs supply (x_t, m_t, ℓ_t) references. Faults flip final/first cues (replaying history) or remove the clean preferred action. Rank absolute individual responses; abstain below magnitude 0.02 or top-two margin 0.01. Clean effects are zero by construction, so perfect gating does not validate retrieval.

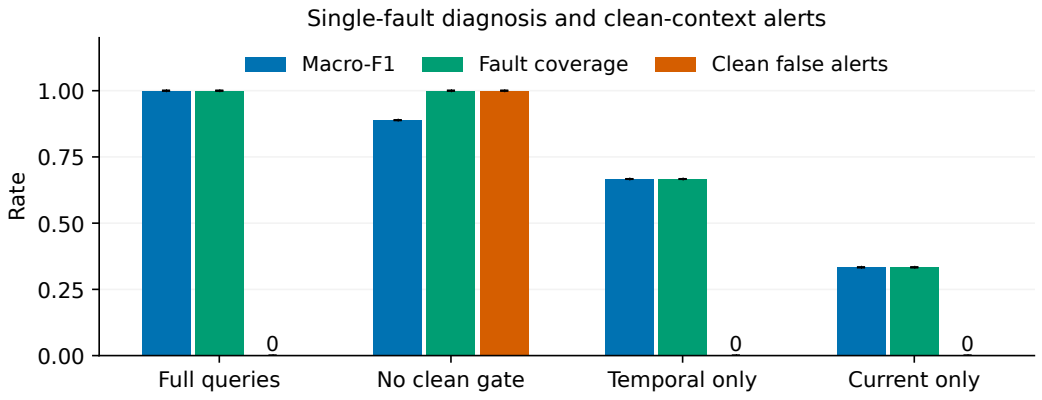

Figure 14: Single-fault query and gate ablations, redrawn from five-seed records. Clean alerts are undesirable, whereas larger F1 and coverage are desirable. Error bars are seed standard deviations (zero to displayed precision); numerical values are retained in Table 32. These are query-availability ablations, not TimeSHAP/WinIT comparisons.

Per seed, 400 episodes generate three faults plus clean contexts (8,000 total). Nonfinite observation/memory and empty-mask edits on the first 50 episodes yield 750 rejected queries; other semantic/version faults are untested.

Robustness/restoration use 300/200 base episodes per seed. Cross seven fault sets (clean, three singles, three pairs) with six scenarios: strengths $s \in \{0.25, 0.5, 1\}$ at zero noise, then $s = 1$ with reference-noise SD 0.05, 0.15, 0.30. Current cues become $(1 - 2s)x$ and memory $(1 - s)m + sm_{\text{opposite}}$; mask faults stay hard. Gaussian noise changes reference observation/memory, not legality. Interpolated states need not be attainable by replay; this is a sensitivity stress test.

Injected-set micro-F1 and active-effect micro-F1 ($|\Delta_j| \geq 0.02$ under clean references) have different targets. Noisy-reference predictions yield active F1=1 by construction at zero noise; SD 0.30 gives F1=0.872, clean alerts=0.055. Error bars span five seed means.

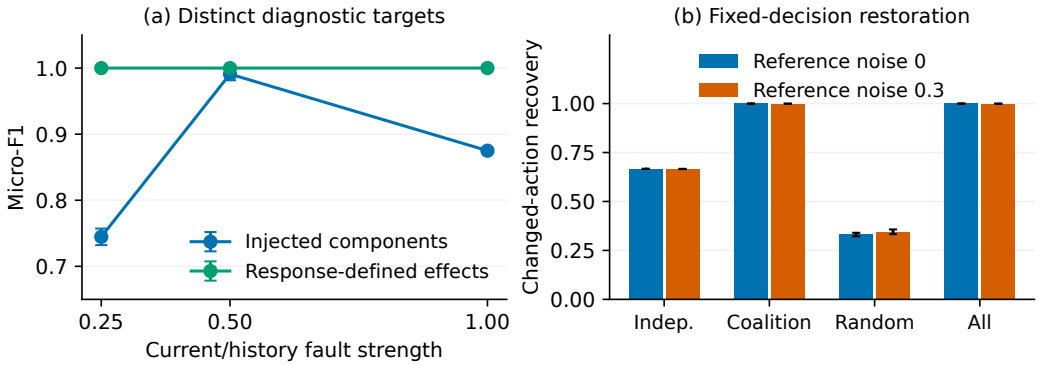

Figure 15: Supplied controlled diagnostic results, redrawn from per-seed measurements. (a) Injected components versus response-defined effects; the latter's perfect clean-reference score follows its definition. (b) Recovery on action-changing fault contexts. Bars/lines are means and error bars population standard deviations over five seeds, not confidence intervals over repeated contexts.

M.5 Reference-Guided Single-Decision Restoration

For reference distribution p_b and replacement distribution p_S, search all eight component subsets for minimum $|S|$ with $\frac{1}{2}\|p_S - p_b\|_1 \leq 0.02$, breaking ties by distance then component names. Full reference replacement guarantees feasibility, not true-clean agreement under noise. Independent replacement selects individual responses ≥ 0.02 in magnitude; random selects the coalition method's count; repair-all replaces three components. All share the reference. Random matches replacement count, not inference cost.

Recovery conditions on changed clean argmax; costs average six nonempty fault sets. All 6,000 full-strength/noise-free faults change actions. Clean cases are scored separately.

Table 25: Full-strength restoration: five-seed mean ± population standard deviation. Recovery is conditional on an action-changing fault; unnecessary repair is any nonempty replacement in a clean context.

Reference noise	Method	Recovery	Components	Clean repair
0	Independent	0.667 ± 0.000	1.167 ± 0.000	0.000 ± 0.000
0	Coalition	1.000 ± 0.000	1.500 ± 0.000	0.000 ± 0.000
0	Matched random	0.331 ± 0.009	1.500 ± 0.000	0.000 ± 0.000
0	Repair all	1.000 ± 0.000	3.000 ± 0.000	1.000 ± 0.000
0.30	Independent	0.666 ± 0.001	1.485 ± 0.004	0.049 ± 0.009
0.30	Coalition	0.999 ± 0.001	1.529 ± 0.002	0.058 ± 0.010
0.30	Matched random	0.345 ± 0.012	1.529 ± 0.002	0.058 ± 0.010
0.30	Repair all	0.999 ± 0.001	3.000 ± 0.000	1.000 ± 0.000

At noise 0.30, clean-action preservation is 99.9% despite 5.8% clean edits. Environment-return benefit and larger searches are untested.

Restoration controls. Figure 17 retains all per-fault results (200 episodes/seed/set). Independent edits fail on CK/HK because the other fault masks the benefit of relaxing legality.

The injected-set oracle shares the reference: recovery is 1.000 (exact) or 0.9995 ± 0.0007 (noise 0.30), cost 1.500, no clean edits. Its information advantage makes it neither a deployment baseline nor a universal noisy-reference bound.

N Extended Related Work and Method Roles

SHAP allocates a model-output difference over feature coalitions; Integrated Gradients integrates a gradient along a reference path (Lundberg & Lee, 2017; Sundararajan et al., 2017). A leave-one-out (LOO) response instead measures removal from the full input, so interactions and redundancy can separate its ranking and support from Shapley allocation. For sequential models, TimeSHAP offers feature-, event-, and cell-level explanations with sequence pruning (Bento et al., 2021); WinIT summarizes distributional effects over prediction windows using conditional replacement (Leung et al., 2023). FIT scores changes in predictive distributions (Tonekaboni et al., 2020); dynamic masks and contrastive perturbations provide other temporal estimators (Crabbé & van der Schaar, 2021; Liu et al., 2024). TSSA instead attributes distribution-shift performance degradation to temporal data properties (Liu et al., 2025). Their interpretation depends on the modeled computation, output functional, and replacement or path; coalition estimators additionally define players. Our protocol specifies those choices for a recorded policy decision; it can host temporal attribution rather than replace it. Section 5.1 runs the public TimeSHAP implementation with a matched event game.

Table 1 in the main text distinguishes estimator roles from the query contract.

RL explanations include recurrent-agent perturbations, structural causal accounts, interpretable policies, reward analysis, and planning representations (Greydanus et al., 2018; Madumal et al., 2020; Wabartha & Pineau, 2024; Ye et al., 2025; Liu & Zhu, 2025; Bush et al., 2025). Other work targets future outcomes, behavioral segments, multi-objective trajectories, or textual explanations (Ruggeri et al., 2025; Rishav et al., 2025; Rajapakse et al., 2026; Terra et al., 2026). EvalXRL proposes evaluation through bug diagnosis and repair as a planned benchmark (Rachum et al., 2026). We study the conditional response of one recorded decision. Saliency sanity checks and faithfulness studies motivate verifying the explained computation (Adebayo et al., 2018; Yeh et al., 2019; Zhang et al., 2024; Makelov et al., 2024; Zheng et al., 2025). Action masking changes that computation independently of sensory content (Huang & Ontañón, 2022). Our responses concern policy computation, whereas counterfactual Shapley credit assignment also separates skill from environmental luck (Li et al., 2026).

O Independent-Reference Study: Design and Full Results

Splits and retrieval. Reuse all six 8/64-step checkpoints. Each checkpoint uses 256 bank episodes, 128 calibration episodes, and 128 test episodes. RNG domains separate date, study, seed, horizon, and split from training. Retrieve within mode by intermediate-distractor Hamming distance, ties by bank ID; exclude endpoints, clean counterparts, labels. Independently drawn banks contain 40/46/54 eight-step test vectors, none at 64; coincidence exposes no counterpart ID.

Faults and units. Current/history faults flip or zero final/initial cues; synthetic post-logit masks remove the task-correct action. Six fault sets (C,H,K,CH,CK,HK), clean contexts, and two edit families yield 10,752 contexts/53,760 method rows from 768 episodes. Active means observed/clean TV > 0.05; recovery means restored/clean TV ≤ 0.05. Clean modification is any replacement; damage changes argmax (coinciding only in these sign-flip results).

Calibration and search. Freeze gates at clean-calibration order statistic 123/128 (95% quantile). Thresholds 0.999822–0.999999 reflect weak retrieval; abstention does not certify cleanliness. Stop if the empty set meets TV 0.05; otherwise exhaustive search minimizes count then discrepancy over

eight subsets, while greedy adds minimum-discrepancy components. Both permit eight queries/four unique sequences, without input hierarchy.

Table 26: Per-checkpoint sign-flip results. Each row has 128 test episodes and 768 injected-fault contexts; only the listed active contexts enter recovery. Recovery is averaged within episode, then across episodes. Exact and greedy searches coincide in recovery and replacement count.

Steps	Seed	Active n	Recovery	Gated recovery	Clean edits	Gated edits
8	11	640	0.492	0.103	0.508	0.016
8	29	593	0.555	0.204	0.445	0.023
8	47	640	0.500	0.186	0.500	0.078
64	11	640	0.523	0.181	0.477	0.094
64	29	550	0.508	0.237	0.492	0.023
64	47	640	0.523	0.092	0.477	0.039

Reference/clean action mismatch is 50.8/44.5/50.0% (8 steps), 47.7/49.2/47.7% (64), distinct from label mismatch. Ideal clean-pair recovery is 1 by information advantage.

Uncertainty and the second edit family. Intervals use 2,000 paired base-episode bootstrap draws, retaining fault clusters; average within episode, then across seeds. Inference conditions on checkpoints. Exact-minus-greedy recovery/replacement-count intervals are $[0, 0]$ for these records, not population equivalence.

Table 27: Paired episode-bootstrap changes from enabling the calibrated gate in exhaustive search, sign-flip family. Negative recovery differences are a cost of fewer clean edits.

Steps	Seed	Recovery change [95% interval]	Clean-edit change [95% interval]
8	11	-0.389 $[-0.470, -0.311]$	-0.492 $[-0.578, -0.406]$
8	29	-0.351 $[-0.424, -0.281]$	-0.422 $[-0.508, -0.336]$
8	47	-0.314 $[-0.383, -0.244]$	-0.422 $[-0.508, -0.344]$
64	11	-0.342 $[-0.417, -0.272]$	-0.383 $[-0.469, -0.305]$
64	29	-0.270 $[-0.330, -0.212]$	-0.469 $[-0.555, -0.383]$
64	47	-0.431 $[-0.516, -0.348]$	-0.438 $[-0.531, -0.344]$

Table 28: Separate zero-erasure fault family; zero cues are outside the binary training support. The calibration threshold is unchanged. Clean results use the same clean episodes.

Steps	Seed	Active n	Recovery	Gated recovery	Gated clean edits
8	11	555	0.492	0.137	0.016
8	29	596	0.555	0.198	0.023
8	47	584	0.500	0.200	0.078
64	11	572	0.523	0.194	0.094
64	29	550	0.508	0.216	0.023
64	47	581	0.523	0.105	0.039

Held-out quality and reference stability. Rank absolute zero-replacement responses at the factual action; evaluate the chosen endpoint by held-out sign flip. A random score permutation matches four queries (factual, two generation, one evaluation). Ties within 10^{-6} use an independent seeded draw; random selection itself does not require generation calls.

Four retrieved references define exact three-component replacement Shapley values at the same target, separate from the input hierarchy. Values $\leq 10^{-6}$ in magnitude are zero; absolute values rounded to six decimals receive average tie ranks. Correlations exclude entirely tied vectors, with counts retained. Active sign agreement includes entries nonzero under either reference (zero/nonzero disagrees), unlike the public task's common-nonzero definition; rates are not pooled.

Table 29: Held-out sign-flip effects and paired episode-bootstrap intervals. Absolute fixed-action responses are in probability units.

Steps	Seed	Ranked	Randomized	Difference [95% interval]	Ties
8	11	0.914	0.500	0.414 [0.312, 0.508]	0.156
8	29	0.851	0.539	0.312 [0.234, 0.398]	0.000
8	47	1.000	0.531	0.469 [0.391, 0.555]	0.000
64	11	0.781	0.539	0.242 [0.133, 0.352]	0.141
64	29	1.000	0.492	0.508 [0.422, 0.594]	0.000
64	47	0.859	0.445	0.414 [0.304, 0.516]	0.172

Table 30: Sign-flip family, reference sensitivity over four retrieved references. Each seed has 896 contexts (including clean cases).

Steps	Seed	Rank correlation	All-sign agreement	Active signs	Zero fraction
8	11	0.597	0.676	0.456	0.503
8	29	0.600	0.654	0.460	0.454
8	47	0.631	0.675	0.473	0.477
64	11	0.695	0.686	0.440	0.536
64	29	0.594	0.642	0.424	0.483
64	47	0.630	0.642	0.433	0.484

Policy-query and physical execution costs. Queries return masked distributions. Memoization limits a context to four sequences; mask variants reuse logits, with no cross-context discount. Offline bank/calibration costs are 256/128 sequences/checkpoint; retrieval needs no neural calls. Active-fault exact/greedy sequence counts average 1.766/1.597 (8 steps), 2.130/1.824 (64), distinct from Table 4's query counts.

The offline audit evaluates four references, ideal pairs, and held-out outcomes irrespective of online requests. Total sequence counts (bank/calibration/clean oracle/held-out/retrieved/ideal) are 1,536/768/768/3,072/132,039/21,504; vectorized calls are 6/6/6/768/43,008/10,752. Outputs, byte keys, calibration, and hashes are retained; full-reference error is at most 1.19×10^{-7}.

P Native POPGym Experiment and Reconstruction-Check Ablation

Environment and training. Unmodified `popgym==1.0.7 RepeatPreviousEasy` (Morad et al., 2023) draws four categories from a 52-card deck with four actions. Stock $k = 4$ rewards $a_t = x_{t-3}$ after three unrewarded decisions: 51 decisions/48 rewards of $\pm 1/48$, maximum return 1. Episode length differs from retention horizon. Native observations/rewards/seeds are retained.

A 32-unit GRU maps one-hot categories to four logits. Seeds 11/29/47 use Adam 0.003, gradient cap 1, entropy 0.01, batch 64, 900 updates, retaining all final checkpoints. Rewards are scaled by 48 with per-step batch-mean baseline; no supervised loss. Action-independent dynamics permit same-seed collection/action passes with observation equality checked. Immediate rewards suffice; each seed has 2,937,600 sampled actions plus collection.

Evaluate 256 new episodes/seed against same-deck uniform controls; query 64 further episodes at $t = 12, 24, 36, 48$, with disjoint seeds and factual argmax. These are newly trained policies. A serialization fix reused the saved checkpoint without retraining/seed changes.

Queries, references, and native legality. Players $t, t-1, t-2, t-3$ replace missing category c by $(c + s) \bmod 4$, $s = 1, 2, 3$: 16 coalitions/game, 2,304 games/36,864 outputs. Edits may violate deck inventory. Current queries retain cache; history replays from before the edit. Maximum replay/efficiency errors are $1.19 \times 10^{-7}/5.23 \times 10^{-8}$.

All-legal native masks give identity constraint queries; removing the selected action is synthetic. Categorical events have no entity hierarchy, so temporal Shapley cannot validate the optional input hierarchy.

Held-out intervention and uncertainty. Shift-one Shapley selects maximum absolute contribution; evaluate with shift two or predecessor replacement, retaining 22.3–23.0% no-ops. Uniform/current controls average four events/select current. Budgets differ: attribution uses 16 outputs, fixed selection none.

Use 2,000 bootstrap draws of 64 episodes with four targets clustered, conditional on checkpoint. All seeds have held-out accuracy/return 1 (float32 return 0.99999994); degenerate intervals are not population guarantees.

Table 31: All public-task seeds. Freeze response uses the held-out predecessor replacement; the last column is selected-minus-current response with a paired episode-bootstrap interval. Native random returns use the same decks.

Seed	Return	Random return	Selected freeze	Uniform freeze	Selected–current [95% interval]
11	1.000	-0.489	0.769	0.192	0.769 [0.714, 0.820]
29	1.000	-0.486	0.777	0.194	0.777 [0.734, 0.816]
47	1.000	-0.489	0.777	0.194	0.777 [0.722, 0.832]

Reference stability. Lag three ranks first under every shift. Between shifts one/two, mean whole-ranking correlations are 0.722/0.713/0.691 after 10^{-6} rounding and average tie ranks. Common-nonzero sign agreement is 0.710/0.750/0.740; no context is entirely tied or lacks common nonzero entries. Near-zero fractions across three references are 1.30/0.85/0.55%. Nondominant mean absolute values $(1.20, 1.35, 1.54) \times 10^{-4}$ explain why minor-sign instability need not undermine the dominant dependency. Raw untoleranced metrics are retained; these summaries concern the specified replacements.

Same-query, same-computation ablation. Inject omission of the final prefix observation: 256 clean/256 faulty contexts per seed. Both gate settings compute six factual/current/history cached/full outputs plus two prefixes (eight invocations). Reject vector discrepancy $> 10^{-5}$; response error $> 10^{-4}$ is misleading. Thresholds are fixed.

All clean cases pass and omission faults fail. On 768/768 balanced contexts, coverage/risk changes from 1/0.5 to 0.5/0; fault-stratum risk is undefined after rejection. This tests screening, not natural failure, repair, or speedup. Independent replay matches saved probabilities/rewards/observations exactly.

Provenance and limitations. Retained assets include environment source/license/hash, three checkpoints, traces/curves/coalitions/logs, and benchmark paper/metadata. Unknown-reference repair, long attention, industrial legality, and environment intervention benefit remain untested.

Q Complete Paired-Reference Diagnostic Results

These single-decision paired-reference results use Appendix M's five-seed supervised GRU and known clean pairs. Independent reference discovery is separate (Section 5.3).

Query and gate ablations. Table 32 covers 6,000 faults/2,000 clean contexts; all 750 invalid queries are rejected. These are query/gate ablations.

Table 32: Supplied single-fault study: means over five seeds. All displayed population standard deviations are zero to shown precision. Coverage and selective accuracy use fault contexts; macro-F1 includes clean false positives.

Queries / gate	Macro-F1	Coverage	Selective accuracy	Clean alerts
Three / enabled	1.000	1.000	1.000	0.000
Three / disabled	0.889	1.000	1.000	1.000
Temporal / enabled	0.667	0.667	1.000	0.000
Current / enabled	0.333	0.333	1.000	0.000

Reference sensitivity. Robustness and restoration reuse 300 and 200 episodes per seed over six scenarios and seven fault sets (63,000 and 42,000 contexts). At strengths 0.25, 0.5, 1, injected-set F1 is 0.745, 0.991, 0.875; Figure 16 separates active-effect F1 and noise.

Full reference replacement guarantees feasibility. Random controls match replacement count, not calls; the oracle knows faults. Outcomes concern single decisions, not environment/industrial recovery.

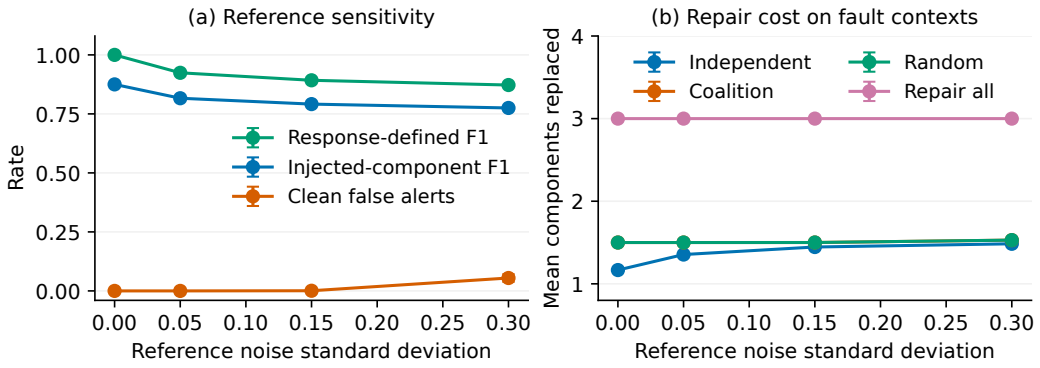

Figure 16: Reference sensitivity at full current/history fault strength. Error bars show population standard deviations over five seeds. Left: separately defined fault targets and clean alerts. Right: replacement count; it excludes policy-evaluation cost.

Diagnostic reference construction. For context (x_t, m_t, ℓ_t), a clean pair supplies all three references. Faults flip the final cue, replay a flipped first cue, or remove the clean preferred action. Rank absolute fixed-target responses to individual reference replacements; abstain below magnitude 0.02 or top-two margin 0.01. Clean-pair effects are identically zero on clean cases, explaining the perfect gate without validating reference retrieval.

Sensitivity protocol and targets. Robustness/restoration use 300/200 base episodes per seed. Cross seven fault sets (clean, three singles, three pairs) with six scenarios: strengths $s \in \{0.25, 0.5, 1\}$ at zero noise, then $s = 1$ with reference-noise SD 0.05, 0.15, 0.30. Current cues become $(1 - 2s)x$ and memory $(1 - s)m + sm_{\text{opposite}}$; mask faults stay hard. Gaussian noise changes reference observation/memory, not legality. Interpolated states need not be attainable by replay; this is a sensitivity stress test.

Injected-component micro-F1 targets the injected set; active-effect micro-F1 targets clean-reference $|\Delta_j| \geq 0.02$, with predictions using the noisy reference. At zero noise, active-effect F1=1 by construction; noise 0.30 gives 0.872 and clean alerts 0.055 (Figure 16). Error bars summarize five seed means, not repeated-context significance.

Recovery conditions on faults changing the clean argmax; cost averages all six nonempty fault sets. All 6,000 full-strength, noise-free fault contexts change the action. Clean cases are separate. Per-fault results in Figure 17 explain independent replacement's failures on mask-containing pairs.

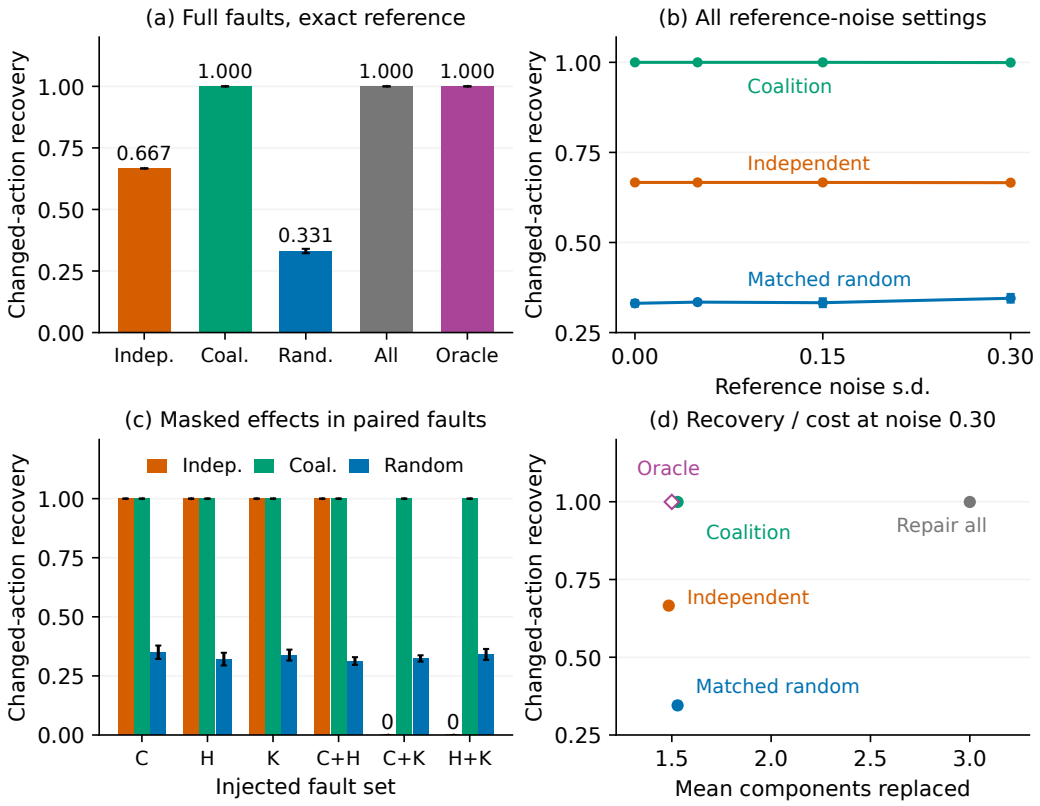

Figure 17: Complete single-decision restoration comparisons from the same supplied records. (a) Exact-reference recovery, including the injected-set oracle. (b) All four reference-noise settings at full strength. (c) Per-fault recovery: C=current, H=history, K=constraint; individual zero bars are labeled. (d) Noisy-reference recovery versus replacement count; the hollow diamond marks the oracle. Error bars are population standard deviations across five seeds. Random replacement matches the coalition method's replacement count, not its inference cost.

Joint restoration and controls. For reference distribution p_b and replacement distribution p_S, search all eight component subsets for minimum $|S|$ with $\frac{1}{2}\|p_S - p_b\|_1 \leq 0.02$, breaking ties by distance then component names. Full reference replacement guarantees feasibility, not true-clean agreement under noise. Independent replacement selects individual responses ≥ 0.02 in magnitude; random selects the coalition method's count; repair-all replaces three components. All share the reference. Random matches replacement count, not inference cost.

At full strength with exact references (Figure 17), independent recovery is 1 for C, H, K, CH and 0 for CK, HK: the other fault hides the benefit of individually relaxing legality. Coalition recovery is 1 throughout; matched random gives 0.350, 0.321, 0.338, 0.313, 0.324, 0.341. Each set uses 200 base episodes per seed across all five seeds.

The oracle knows injected components but uses the same reference: exact-reference recovery 1.000 at cost 1.500; noise-0.30 recovery 0.9995 ± 0.0007 at cost 1.500 with no clean replacements. This information-advantaged control is neither a deployment baseline nor a universal bound under noisy references.

At noise 0.30, coalition restoration preserves 99.9% of clean actions while editing 5.8% of clean contexts. Neither environment-return gains nor larger-subset scalability is tested.