MEASURING THE LOCAL ACTION UTILITY OF VISUAL HISTORY IN DREAMER

Anonymous authors
Paper under double-blind review

ABSTRACT

An agent can see the same image yet choose a different action because of its visual history. Does that change help it succeed? We measure this link by holding the current image fixed, changing past images, and evaluating the resulting first-action probabilities with the same continuing agent. A second measurement asks whether different current cues lead to different later outcomes when the first action is held fixed. Both use real simulator continuations and an explicit pairing of the recurrent policy's stochastic states. Controlled tasks distinguish information that is stored, information that changes actions, and information that improves outcomes. In Craftax-Classic, an exploratory DreamerV3 cohort motivates a prespecified confirmation on five new training seeds. Recent history improves local target success through the first-action choice by 5.13 and 5.36 percentage points under two masked continuation conditions. Every new model has a positive estimate, and joint model/world intervals remain positive after adjustment for three primary endpoints. An exploratory analysis links this gain to a shift toward the nominated tree; any-wood and native-reward gains remain unresolved. The separate continuation-square endpoint is also unresolved. The confirmed benefit concerns recent history with a visible current target, erased-history recipients, and masked later observations. A full-panel extension with clean recipient history and unmasked future images leaves its benefit unresolved. The measurements connect visual input to local action utility under explicit execution conditions.

1 INTRODUCTION

Consider an agent approaching a tree to collect wood. The tree is visible in its current image, but the preceding images may still influence its next move. If editing those past images reduces its chance of collecting wood, what has changed? History may help the agent choose a useful first move, or it may guide what the agent does afterward. An outcome difference between two freely acting agents combines these possibilities.

This distinction matters when studying memory in a world model. Information about a past image can be present in the recurrent state without affecting a decision. A decision can depend on that information without improving the outcome. To understand its behavioral value, we need to connect the input, the action it changes, and the consequences of that action. We study these connections in trained DreamerV3 agents acting in Craftax-Classic (Hafner et al., 2023; Matthews et al., 2024).

Our first measurement asks whether history makes the next action more useful. We replay the same recorded actions with clean or edited past images, keep the current image identical, and read the two action distributions. Each distribution is then evaluated using the same set of simulator continuations. The source supplies only the first-action probabilities; the agent that continues keeps its own recurrent state. This isolates the benefit of the changed action choice under a specified continuation. We call it *history-source gain*, G.

Our second measurement asks about behavior after the first action. In a separate intervention, two branches receive different current cues after the same edited history. We force both to take the same first action and compare their subsequent success probabilities. Squared differences capture local responses that a signed average can hide; we call this measurement the *continuation square*, D_w. It changes a different input at a different time from G, so the two measurements are not additive parts

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.

of one memory effect. For stochastic recurrent states, their definition also specifies which random inputs are shared between branches.

The strongest empirical result concerns the first action. An exploratory five-model cohort motivates a prespecified confirmation on five new training seeds. The confirmation finds history-source gains of 5.13 and 5.36 percentage points under two continuation conditions, positive in all five models and under joint model/world intervals adjusted for the three primary endpoints. The tree remains visible at the decision, and nearly all anchors include a visible exposure one frame earlier. Thus even this recent history has local action value under the tested conditions. The independent continuation-square estimate remains uncertain.

The contribution is a way to measure the outcome value of a history-dependent action choice separately from cue-dependent behavior after that choice. Exact examples and trained policies with known information paths show why these questions differ from whether past information is decodable or affects a gradient. A stochastic-state calibration further checks the squared estimator under several specified pairings. The Dreamer study then supplies independent evidence of local history-source benefit. It leaves open whether the measurements can guide an intervention that improves complete-task performance; Section 6 develops that next step.

2 Following information into actions and outcomes

Consider an agent beside a tree. Past images may help it choose to turn toward the tree. Once that turn is fixed, its internal state may also affect what it does next. We study these two questions with different interventions: G changes past images while preserving the current image; D_w changes the current cue and fixes the first action (Figure 1). They are complementary measurements, not components of one memory effect. Both G endpoints and D_w remain the prespecified primary endpoints of the independent confirmation.

An *anchor* is a recorded simulator fork. At that fork, an action *source* supplies only probabilities for the first action. A *recipient* executes that action and retains its own recurrent state for the rest of the rollout. Keeping these roles separate lets us ask whether history improves action selection under a specified continuation.

2.1 Does history improve the first action?

We replay the recorded actions with three image sequences. The clean source sees the unmodified past and current images. Source T sees past candidate tiles erased and the true current image; source E sees both past and current candidate cues erased. Their first-action distributions are p_{clean}, p_T, and p_E. Thus clean versus T changes history while holding the current image fixed. At our Craftax anchors, that current image already shows the adjacent target tree: the past is not the only source of target information.

To evaluate each action, we construct two recipients from the same erased prefix. Recipient T sees the true current cue once; recipient E sees its erased counterpart. Let Z_c be the resulting recurrent state for $c \in \{T, E\}$. We force first action a and execute 15 further native actions, masking the candidate throughout. The expected real 16-action outcome is $Q_c(X, Z_c, a)$, where X fixes the checkpoint, recorded prefix, simulator fork, and eligibility. Masking prevents reacquisition of candidate pixels except on the protected player tile. These values describe the masked continuation, not naturally visible play.

Keeping state and randomness aligned. Dreamer's actor reads the recurrent state $z_t = f_\theta(z_{t-1}, o_t, a_{t-1}; u_t)$, including the posterior random input u_t; the action executed enters the next update. We share posterior random inputs at matched updates when constructing clean, T, and E states. For Dreamer's categorical posterior this is a Gumbel-max pairing in the learned category order. It asks how the response changes with the images when the sampler's random inputs are held fixed. Each branch still has its own state; T and E share their erased-prefix state before the current image. Write γ_X for the resulting joint law of $(Z_{\text{clean}}, Z_T, Z_E)$, and abbreviate $Q_c(X, Z_c, a)$ to $Q_c(a)$.

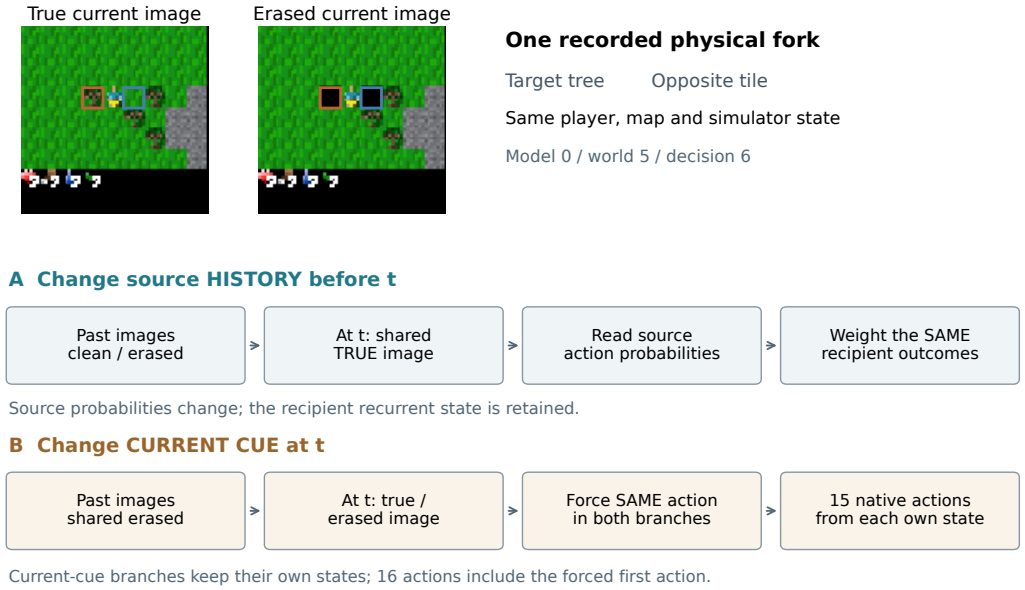

Figure 1: Two intervention times at a recorded Craftax fork. A changes past source images while keeping the current source image true; only the resulting action probabilities weight the continuation outcomes. B changes the current cue after a shared erased prefix, then forces the same first action in both branches. Source states never replace continuation states. These complementary measurements are not an additive decomposition of one history effect. Images are recorded pixels and their candidate-tile erasure.

For any shared source distribution s, the recipient's expected outcomes are

$$V_c^{\text{free}} = p_c^\top Q_c, \qquad V_c^{\text{fixed(s)}} = s^\top Q_c. \tag{1}$$

Free execution lets the recipient choose its first action; fixed execution uses the supplied source probabilities. A single forced action is the special case $s = \delta_{a^*}$. In neither case does the source replace the recipient's state.

The history-source gain asks whether clean history places more probability on actions with better outcomes:

$$G_c^\gamma = \mathbb{E}_{X,\, Z\sim\gamma_X}[(p_{\text{clean}} - p_T)^\top Q_c]. \tag{2}$$

For an illustrative two-action example, suppose the recipient succeeds with probability 0.6 after turning toward the tree and 0.2 after turning away. Moving 0.1 probability from the second action to the first gives $0.1(0.6 - 0.2) = 0.04$, a four-percentage-point gain. The continuing agent is unchanged; only the chances of choosing its first action change. We evaluate the source shift separately with T and E recipients. The different contrast $(p_T - p_E)^\top Q_c$ would change the current source image instead. The p_T comparison was added after the original two-source study.

We enumerate all 17 first actions and estimate their outcomes with paired real simulator continuations, rather than a learned critic. Each product is computed within a posterior draw before averaging over draws, anchors, and models, preserving dependence between action probabilities and recipient states. Consequently, G measures the benefit of supplying first-action probabilities to a retained recipient, not the return improvement of a complete agent given clean history.

2.2 Does a cue change what happens after the same action?

For the second question, both recipients take the same first action. Their expected outcome difference is $\delta(a) = Q_T(a) - Q_E(a)$. A signed average can hide differences that help in some states and hurt in others. We therefore measure their squared size:

$$D_w^\gamma = \mathbb{E}_{X,\, Z\sim\gamma_X}\left[\sum_a w(a \mid X, Z)\delta(a)^2\right]. \tag{3}$$

The nonnegative weights sum to one and are chosen before tail execution; we use $w = (p_T + p_E)/2$. They may depend on the sampled states. We write D_w for the original pairing. It characterizes that pairing: changing γ_X can change D_w even if each branch's marginal distribution stays the same. Averaging each branch's values over posterior draws before squaring answers a different question (Appendix D).

Squaring a noisy rollout difference would also count sampling noise as a response. Instead, we run two paired tail repeats, form $d_r(a) = Y_{T,r}(a) - Y_{E,r}(a)$ from observed outcomes, and average $\sum_a w(a)d_1(a)d_2(a)$. If the repeat pairs are independent conditional on X, the posterior tuple, and a, this cross-product estimates the desired square without additive tail-noise bias. T and E may share randomness within a repeat. Finite estimates can still be negative, and we retain them. Appendices B and D give the derivation and alternative pairing analyses.

2.3 Which decisions and outcomes do we measure?

Anchors are selected before branching. Cue edits change rendered candidate tiles, and masks track fixed world coordinates while preserving player-occupied tiles. Every attempted world is logged, absent anchors are not replaced, and failures and early deaths remain outcomes.

Target success requires the nominated tree to supply the first wood and survival through the horizon. Target-directed first turn, any-wood success, and environment-reward return are separate endpoints. Returns use the original rewards under masked observations, rather than full-task native play.

We average draws within anchors, eligible anchors within models, and the five models equally. Ten thousand paired reset-world resamples preserve cross-model dependence and recalculate eligible denominators. World-only intervals condition on the five checkpoints; the independent confirmation also jointly resamples training models and worlds and adjusts for its three primary endpoints (Section 5.1). Models can reach different anchors in the same reset world. Common-state comparisons and training-seed sensitivity are reported separately.

Appendix N defines the supporting free-action history contrast B_{history} and the reference-dependent cue interaction I_s. They answer additional questions, rather than completing a single memory score.

3 Calibration with known information paths

Before applying the measurements to Dreamer, we test what they distinguish in a small task where we control the available information paths. An agent sees a past target cue x, a blank frame, and a current cue that reveals x in half the states. It chooses route a, then sees that route and chooses b. An unobserved fair coin rewards either $a = x$ or $b = x$.

The visible route creates two ways to succeed. The agent can remember x for both decisions, or use its memory to choose the route and then follow that visible route. We train three policies with the same architecture but different carry rules: reset every frame, reset only after a, or never reset. The second rule permits history to influence the first action while preventing recurrent information from surviving it.

3.1 Separating action choice from later behavior

The four-frame, two-action task has three carry rules, two observability settings, and ten initializations, giving 60 trained policies. The main comparison uses the 30 partially observed policies; fully observed twins supply the target throughout. Policies learn by expected-reward optimization with 25% erased-history augmentation. We retain every seed (Appendix H).

The reset-every-frame policy cannot retain the past cue. Both history gain and the continuation square are zero (Table 1). This gives a reference against which to interpret the two memory-capable policies. Additional interaction contrasts and exact counterexamples appear in Appendices H and I.

Resetting after the first action gives a 25-point history gain and zero D_w. History improves the route choice, and the route carries the needed information thereafter. With recurrent carry, history gain is 12.51 points and $D_w = 0.06246$: fixing the route still leaves a continuation difference. All ten recurrent seeds have positive exact D_w; both reset controls have exactly zero. Both memory-capable

Table 1: Known information paths give different measurement profiles, despite similar return in the two memory-capable policies. Means over ten seeds; history gain uses Q_T, with the same values under Q_E at this precision. Return is expected binary success; D_w uses probability-squared units.

Information path	Return (%)	History gain G (pp)	D_w exact
Reset each frame	74.992	0.00	0.00000
Reset after action	99.999	25.00	0.00000
Recurrent carry	99.998	12.51	0.06246

paths achieve nearly 100% return. Their different G values reflect how the recipient behaves when history is erased, illustrating why G must be interpreted relative to that recipient.

3.2 From sensitivity to outcome benefit

Existing tools capture useful distinctions here. POPGym Arcade's observability gap is nearly zero for both successful paths; historical recall density at the first action is positive for both, and final-action gradients also distinguish the recurrent path (Wang et al., 2026). Our measurements add the size of the resulting outcome benefit.

For example, when the current cue is visible, an exploratory breakdown gives historical recall shares of 32.26% and 46.56% for the two successful paths, but history gains of only 0.0008 and 0.0054 points. The current cue already supplies the answer, and performance is saturated. Absolute gradients are tiny even though their normalized historical shares are substantial. When the current cue is absent, history gains are 50.00 and 25.01 points. Thus sensitivity and conditional utility answer different questions; these results do not imply that recall density fails.

A second calibration makes a related distinction between storage and use. The same recurrent architecture retains a perfectly decodable past bit across three reward settings. Flipping that bit's past cue strongly changes final actions when the bit determines reward, but barely changes them when another current or historical bit determines reward. Thus the agent can preserve information that it does not use for the tested decision, even while using other history. Appendix P reports the ten-policy experiment and its payoff and gradient comparisons. These small controlled tasks establish what the measurements distinguish before we apply them to a policy whose learned information paths are unknown.

4 Dreamer at a shared decision point

We study a concrete local decision in Craftax-Classic: an agent is next to a tree, has not yet collected wood, and is not facing the tree. This provides a natural place to ask whether its visual history helps it choose what to do next. We record the first eligible approach in each rollout and restore the same physical state for every comparison. The target endpoint is collecting the first wood from that nominated tree and surviving the 16-action horizon.

The study uses two separate cohorts of five DreamerV3 agents, each trained for roughly one million driver transitions with the same architecture. The actor reads RSSM features and chooses among 17 actions from pixel observations. The original cohort achieves 50.49% mean full-task achievement over 100 native evaluation episodes per model (Appendix J); that evaluation was not repeated for the confirmation cohort. Earlier free-action and common-state experiments motivate the present comparison and are reported separately in Appendices K and Q.3.

Each cohort uses a fresh block of 128 reset worlds per model. The first yields 214 eligible decision points in 89 world IDs; the confirmation yields 215 in 95 world IDs. The target is visible in the true current image. Its most recent eligible visible exposure is one frame earlier at 213/214 and 215/215 anchors, respectively (Appendix O). This is a test of recent history in the presence of a current cue. The 16 subsequent actions measure the consequences of the choice, not a 16-step retention interval.

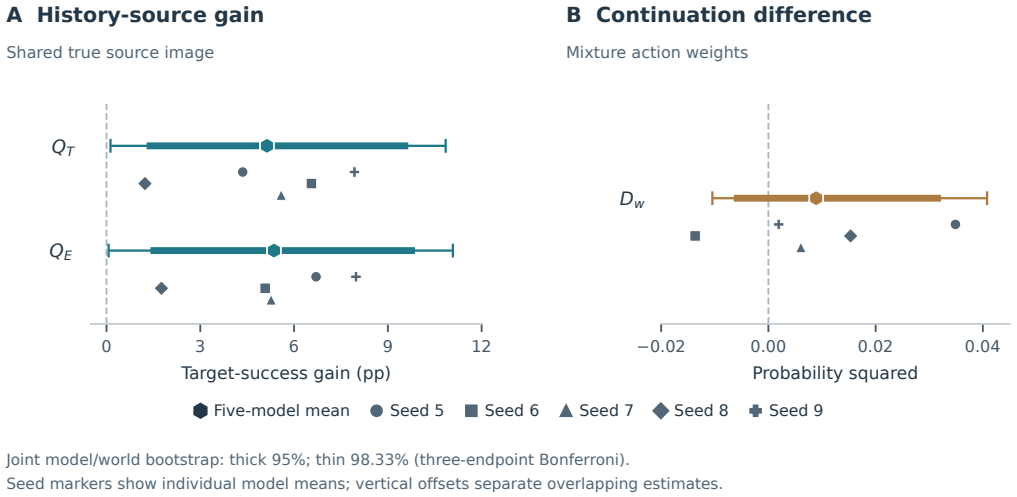

Figure 2: Recent history improves first-action utility in the independent cohort. Small markers show the five new models; hexagons show equally weighted means. Thick bars are joint model/world 95% intervals; thin bars are 98.333% intervals giving a Bonferroni adjustment for three endpoints. Both history gains remain positive after adjustment; the continuation square is unresolved. Vertical offsets separate models; panel units differ.

5 Recent history improves the first-action choice

At each decision, we evaluate every first action from both recipient states. Both recipients start from erased history; they differ in whether their current cue is true or erased. Eight posterior draws and two paired repeats per action give 116,416 simulator rollouts in the original cohort and 116,960 in the confirmation. These outcome matrices let us compare sources while retaining the same continuing agent (Appendix A).

5.1 Independent confirmation on five new training seeds

Before training the new agents, we specified G_T, G_E, and mixture-weighted D_w as three primary endpoints. Every new model and eligible decision enters the analysis. Appendix C records the protocol and recovery from a host-memory interruption.

History improves the action source under both continuation conditions (Figure 2). With a true current cue in the recipient, the gain is +5.13 percentage points, with a joint model/world 95% interval of [1.28, 9.65]. With an erased recipient cue, it is +5.36 [1.41, 9.87]. Both sources see the same true current image, so these gains concern how their different histories change the first action. Every model has a positive estimate in both conditions. The three-endpoint adjusted intervals are [0.13, 10.85] and [0.07, 11.08] points; their near-zero lower bounds reflect limited precision with five training runs. Model-t sensitivity gives the same qualitative conclusion (Appendix O).

What does the agent do differently? An exploratory analysis of these saved outcomes makes the gain concrete (Figure 3). Clean history adds 11.76 percentage points to the probability of acting toward the nominated tree. Target success rises from 65.06% to 70.19% with the T recipient, and from 63.54% to 68.89% with E. Success here means obtaining the first wood from that tree and surviving the horizon. Gains in any-wood success, also requiring survival, are smaller and unresolved: +0.28 [−1.31, 1.68] and +0.64 [−0.46, 2.01] points. Environment-reward gains also remain unresolved. Because any-wood includes both nominated-tree and other-tree outcomes under the same survival requirement, the resolved target gain should be interpreted as local redistribution toward the nominated tree, not as evidence of more wood collected or higher reward overall (Appendix E).

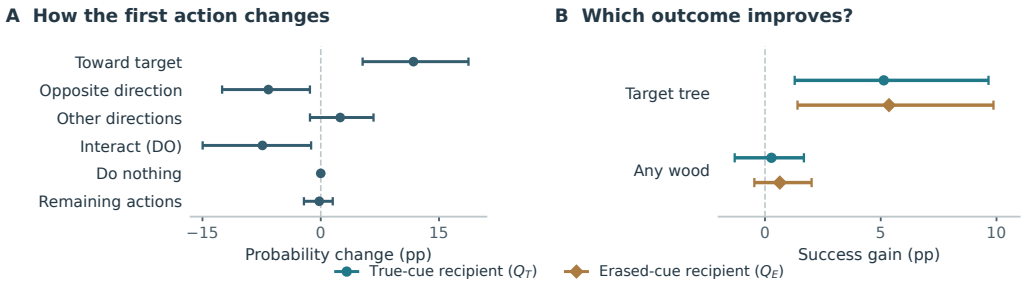

Figure 3: The behavior behind the history-source gain. Left: changes in first-action probabilities, grouped by their relation to the tree. Right: gains in target and any-wood success under each retained recipient. All 215 confirmation anchors are included. Bars are descriptive joint model/world 95% intervals; secondary comparisons are exploratory and unadjusted.

5.2 How the original cohort compares

The original cohort suggested gains of +2.60 and +2.50 points. Its conditional-world 95% intervals, [0.64, 4.77] and [0.47, 4.81], concern those five fixed checkpoints. Later joint model/world resampling gives $[-0.66, 6.42]$ and $[-0.72, 6.19]$ (Appendix O), making confirmation on new training seeds central to the empirical claim. The p_T source and continuation-square analyses were added after the original collection and are exploratory there. We keep the cohorts separate; their differing means do not identify an effect of training or recovery (Appendix B).

5.3 What changes when the first action is held fixed?

The continuation square asks whether current cues change subsequent outcomes after the same first action. In the original cohort, the signed difference is +0.44 points $[-0.45, 1.35]$, whereas $D_w = 0.00712$ with conditional-world 95% interval [0.00205, 0.01300] and exploratory joint model/world interval [0.00033, 0.01578]. A small signed mean therefore need not imply identical continuations, although these summaries do not locate the variation.

The independent cohort leaves this endpoint unresolved: $D_w = 0.00891$, with joint model/world 95% interval $[-0.00645, 0.03218]$. Its units are squared probability, not percentage points. Directly squaring noisy outcome estimates would inflate the original result to 0.02321; weighting by the clean source instead also leaves that original result unresolved (Appendix B).

5.4 What the history edits tell us

Additional sources test whether the gain depends on the edit. In the original cohort, swapping historical candidate crops reduces target success relative to clean history by 19.62 points under Q_T and 18.14 under Q_E. Distant masking at matched area and times gives values 3.48 and 3.53 points above candidate masking (Appendix B.2).

In the new cohort, clean history exceeds candidate grass replacement by 18.92 points under Q_T, with adjusted joint interval [8.54, 32.25]. Nearby black minus candidate black is unresolved at -1.85 $[-9.70, 5.35]$ points. Nearby grass leaves 209/215 histories unchanged, largely repeating the clean comparison. The response thus extends beyond black replacement, but these controls cannot separate loss of useful information from misleading-input harm: grass can suggest false terrain, and edit geometry is imperfectly matched (Appendix R).

5.5 How the execution conditions affect interpretation

Assigning the source to a different posterior draw at the same anchor gives exploratory new-cohort gains of +4.92 and +4.97 points, with joint 95% intervals [1.02, 9.42] and [1.02, 9.46]. Independently pairing T/E draws raises the continuation square to 0.02825, but changes the measured quantity. Averaging over posterior states first also leaves continuation differences unresolved in both cohorts (Appendices D and O).

On a prospectively selected 60-anchor panel, unmasking only the observations after the first action changes G_T/G_E by $-0.62/+0.02$ points. Both recipients still start from erased history. Adjusted change intervals and unmasked-level joint 95% intervals all include zero: the panel establishes neither an unmasked benefit nor equivalence across visibility conditions (Appendix S). Separate process repeats change the original-cohort history gains by less than 0.0005 points (Appendix G).

A clean-recipient extension. Can a recipient with its own clean history also use the changed action source? An extension specified before collection tests all 215 anchors with clean recipient history and unmasked future images. The nominated-tree gain is +3.11 points, with joint 95% interval $[-1.94, 7.70]$ and two-endpoint-adjusted 97.5% interval $[-2.99, 8.63]$. Any-wood gain is -0.53 points and native-return gain is +0.038 reward units; both intervals also span zero. This same-cohort extension leaves clean-recipient benefit unresolved (Appendix F).

6 What the measurements reveal

A visible target need not make its recent history redundant to a recurrent policy. Two histories ending in the same view can produce different action preferences; our measurement asks whether that difference helps. In the independent Dreamer cohort, clean history puts more probability on acting toward the nominated tree and improves local target success by about five points. The recipient retains its own state, so the result concerns the value of that first-action choice. The unresolved any-wood and reward gains limit a broader performance claim.

The primary recipients both begin with erased history. In the full-panel extension, the recipient instead retains clean history and sees unmasked future images. Its target gain is +3.11 points, but target, any-wood, and native-reward intervals all span zero. The evidence therefore leaves benefit under this more natural continuation unresolved. A positive result in one condition and an unresolved result in another do not establish a difference between them; recipient history and future visibility also change together.

The second question remains open: after the first action is fixed, does the current cue change later behavior? The original cohort supports a positive continuation square under the specified weighting and coupling, while the independent cohort leaves it unresolved. This is neither confirmation nor evidence that the effect is absent.

For a stochastic policy, the comparison also depends on how internal randomness is shared. Common Gumbel inputs compare cues under the same sampler inputs; independent inputs additionally include unrelated internal fluctuations. Neither comparison is determined by branch distributions alone. Our known-model calibration recovers analytic targets under shared, independent, and antithetic couplings, including state-dependent weights (Appendix D.6). It validates estimation for a chosen comparison, while that choice remains part of the research question.

The next substantive test is to use the measurement to choose an intervention: identify states in a competent policy where changed information should improve action selection, then validate the prediction on independent states against simpler diagnostics. This needs a reward-aligned outcome, valid information edits, and controlled delays. Our present controls leave useful-information loss versus misleading-input harm unresolved, and the delay study changes visible competence as well as history availability (Appendix Q.9). Restorable simulators with enumerable first actions make these questions executable; broader use needs an execution design suited to the task.

7 Related work and the measurement contribution

Behavioral interventions and memory. Déletang et al. (2021) distinguish internal memory from information carried by an agent's position, using simulator interventions to disrupt a T-maze strategy. Their framework also includes shared exogenous randomness and causal pathway analysis. We build on this perspective by controlling the first-action source and subsequent recurrent execution separately. This makes it possible to ask whether a history-induced action change is useful and whether cue-conditioned outcomes differ after the action is matched.

Pathways and conditional utility. The first-action contrasts connect to direct, indirect, and mediated-interaction effects (VanderWeele, 2014). Recurrent states depend on the exposure, so the simulator interventions and noise coupling must be specified to give this mapping meaning (VanderWeele et al., 2014); Appendix N gives the details. POPGym Arcade studies policy memory through paired observability settings and gradient diagnostics (Wang et al., 2026). We apply these diagnostics to our trained calibrations, where final-action gradients also identify the recurrent path. Historical gradient influence and outcome utility answer complementary questions: an input can affect a decision without improving its consequence.

Memory in world models. Dreamer learns latent dynamics and control (Hafner et al., 2023), while Craftax supports evaluation in an open-ended environment (Matthews et al., 2024). Transformer and state-space world models examine memory-intensive control (Dedieu et al., 2025; Deng et al., 2023; Samsami et al., 2024). Other work combines representation probes with latent interventions (Zhang, 2026), or evaluates structured memory through executed model-predictive control (Lillemark et al., 2026). Object permanence in passive prediction (Liu & Chen, 2026), token correspondence (Kim et al., 2026), historical evidence routing (Yuanruyi et al., 2026), and object- or region-specific blackouts (Barsainyan et al., 2025) examine further aspects of retained context. Our contribution is a complementary way to study fixed trained policies: explicit action-source and continuation measurements, with real-simulator outcomes and estimators that account for stochastic pairing. The independent Dreamer cohort tests this distinction across separately trained policies.

8 Conclusion

Visual history can improve the next action even when the target is still visible. In five independently trained Dreamer agents, retaining recent history improves local target success through the first-action source under two specified continuation conditions. A separate test of cue-dependent behavior after the same first action remains unresolved. Together, the measurements connect visual input to action choice and outcome without assuming that stored information is useful. The confirmed result concerns erased-history recipients with masked later observations. A full-panel clean-history, unmasked extension leaves target, any-wood, and native-reward gains unresolved. Neither a complete-task improvement nor a difference between the execution conditions is established.

Reproducibility statement

The accompanying anonymous archive, `visual-history-measurements.zip`, contains compact per-draw matrices, trained calibration weights, and scripts for the reported recorded-data analyses. It retains all 215 confirmation anchors, including the clean-recipient extension, and documents the study designs and recovery deviation. The archive reconstructs saved statistics and calibration results; it does not rerun Dreamer training or the Craftax simulator experiments.

AI use statement

Generative AI tools were used to aid writing and editing and to draft sections of the paper. The authors reviewed the resulting text and are responsible for all claims, citations, mathematical statements, code, and results in the final manuscript.

References

Aryan Amit Barsainyan, Jing Yu Lim, and Dianbo Liu. STORI: A benchmark and taxonomy for stochastic environments, 2025. URL https://arxiv.org/abs/2509.01793v2.

Antoine Dedieu, Joseph Ortiz, Xinghua Lou, Carter Wendelken, J Swaroop Guntupalli, Wolfgang Lehrach, Miguel Lazaro-Gredilla, and Kevin Patrick Murphy. Improving transformer world models for data-efficient RL. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 12944–12968. PMLR, 2025. URL https://proceedings.mlr.press/v267/dedieu25a.html.

Grégoire Déletang, Jordi Grau-Moya, Miljan Martic, Tim Genewein, Tom McGrath, Vladimir Mikulik, Markus Kunesch, Shane Legg, and Pedro A. Ortega. Causal analysis of agent behavior for AI safety, 2021. URL https://arxiv.org/abs/2103.03938.

Fei Deng, Junyeong Park, and Sungjin Ahn. Facing off world model backbones: RNNs, transformers, and S4. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2307.02064.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models, 2023. URL https://arxiv.org/abs/2301.04104.

Youngin Kim, Ray Sun, Inho Kim, Bumsoo Park, and Hyun Oh Song. Identifiable token correspondence for world models. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2605.16457v3.

Hansen Jin Lillemark, Benhao Huang, Fangneng Zhan, Yilun Du, and Thomas Anderson Keller. Flow equivariant world models: Structured memory for dynamic environments. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2601.01075v2.

Yang Liu and Yuming Chen. Event-conditioned diagnostics of kinematic, contact, and object-permanence structure in passive object-state world models, 2026. URL https://arxiv.org/abs/2606.28455v3.

Michael Matthews, Michael Beukman, Benjamin Ellis, Mikayel Samvelyan, Matthew Thomas Jackson, Samuel Coward, and Jakob Nicolaus Foerster. Craftax: A lightning-fast benchmark for open-ended reinforcement learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 35104–35137. PMLR, 2024. URL https://proceedings.mlr.press/v235/matthews24a.html.

Mohammad Reza Samsami, Artem Zholus, Janarthanan Rajendran, and Sarath Chandar. Mastering memory tasks with world models. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2403.04253.

Tyler J. VanderWeele. A unification of mediation and interaction: A four-way decomposition. *Epidemiology*, 25(5):749–761, 2014. doi: 10.1097/EDE.0000000000000121. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC4220271/.

Tyler J. VanderWeele, Stijn Vansteelandt, and James M. Robins. Effect decomposition in the presence of an exposure-induced mediator-outcome confounder. *Epidemiology*, 25(2):300–306, 2014. doi: 10.1097/EDE.0000000000000034. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC4214081/.

Zekang Wang, Zhe He, Borong Zhang, Edan Toledo, and Steven Morad. Investigating memory in RL with POPGym Arcade. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2503.01450v8. Diagnostic definitions follow arXiv:2503.01450v8.

Yuanruyi, Yue Cao, Haojia Gao, Guanqiu Guo, Ziyuezhang, Shangqin Lin, Junbo Tan, Bokui Chen, Zhuo Zou, and Xueqian Wang. HERA: Historical evidence routing adapter for physical prediction in latent world models, 2026. URL https://arxiv.org/abs/2608.05523v1.

Xinyu Zhang. What do world models learn in RL? probing latent representations in learned environment simulators. In *ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling*, 2026. URL https://arxiv.org/abs/2603.21546.

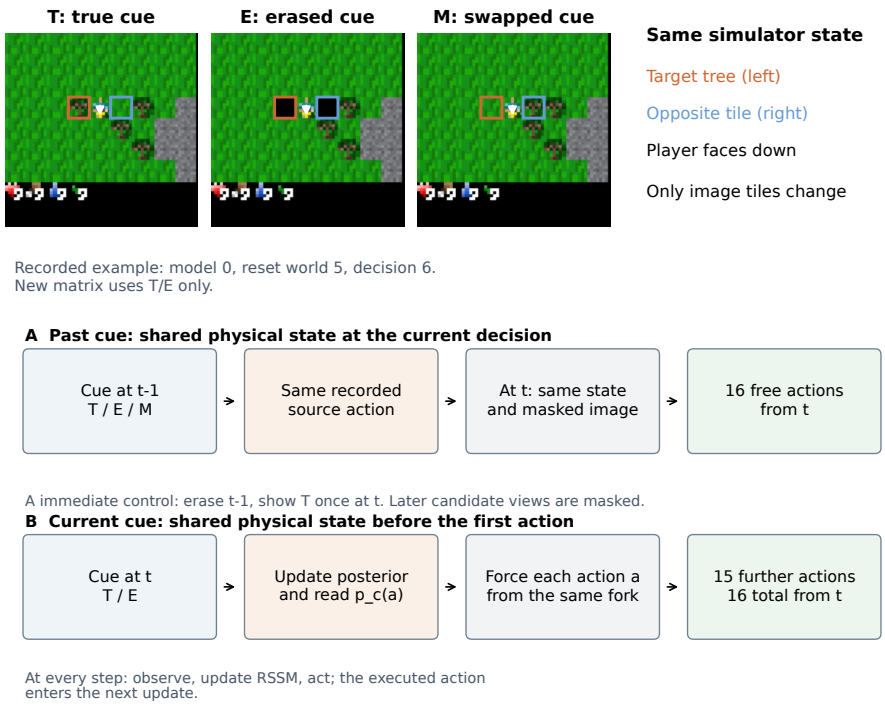

Figure 4: A recorded natural fork and two measurement windows. Top: actual candidate images with unchanged simulator state; T is true, E erased, and M mismatched. M belongs to the earlier content control; the continuation matrix uses T/E. Below: matched-prefix history and cue/action-constraint assays start their 16-action windows at different decisions. Colored outlines identify candidate tiles.

A Continuation-matrix design and additional results

Prospective block. The design specified five existing checkpoints, reset seed 260919, 128 attempted worlds per model, a 128-transition prefix limit, eight posterior draws, two continuation repeats, and a 16-action horizon.

The primary endpoint is nominated-tree success and the primary cue comparison is true minus erased. The sources p_{clean} and p_E and the Q_E-reference decomposition were specified before collection.

Mismatched cues were omitted from this matrix before collection. A separate smoke block at seed 260918 checked execution; its outcomes are excluded. No extra worlds replaced misses or were added after viewing main effects.

Natural eligibility. An anchor has exactly one adjacent tree, no previously collected wood, and at least two earlier candidate exposures without player/entity overlays. The player must not already face the target. The first eligible anchor is used.

Models 0–4 contribute 43, 52, 42, 25, and 52 anchors. These are 214 agent–world pairs over 89 distinct reset-world IDs, not 214 independent world replications. Each model reaches its own physical states.

State and action timing. The T/E branches replay the same erased prefix using actual recorded actions. Each cue frame updates the complete posterior once. Its action probabilities and each forced-action continuation use that same updated state.

Table 2: Per-model interactions on the new natural panels. These model comparisons include differences in reached states; they are distinct from the common-state receiver study.

Model	Anchors	Clean-source I (pp, 95% CI)	Erased-source I (pp, 95% CI)
0	43	+0.82 [−3.43, 5.14]	+0.99 [−2.75, 4.90]
1	52	+1.49 [−4.26, 7.29]	+2.69 [−2.95, 8.57]
2	42	+9.11 [2.91, 16.25]	+7.45 [2.23, 13.71]
3	25	+1.21 [−4.74, 7.49]	+1.96 [−3.98, 8.48]
4	52	−1.25 [−3.56, 0.33]	+0.30 [−0.42, 1.08]

Forcing the first action changes the action executed, not the posterior representation. The actual forced action feeds the next RSSM update. The 15 subsequent steps use the native policy and candidate masking.

The clean source replays clean past images and observes the clean current image at the corresponding random counter. Its distribution weights the T/E matrices; its clean carry never replaces their continuation states.

Products and random pairing. For each posterior draw, two tail repeats estimate the 17 entries of Q_c. Products $p_c^\top Q_c$ and $s^\top Q_c$ are then formed before averaging draws. Averaging p and Q separately would discard their dependence.

Random counters and simulator transition keys are paired across conditions and forced actions. Repeat-specific tail streams separate the two repeats. The first action is included in every 16-action scoring window.

The 116,416 matrix trajectories contain 1,862,605 actual simulator steps. All 23 terminal tails remain in scoring; 21 terminate before action 16. There are also 428 separate sampled-action checks, additional to the matrix trajectory count.

Joint world uncertainty. Each bootstrap replicate samples the 128 reset-world indices with replacement, jointly across all five models. Eligible multiplicities reweight each model's anchor mean; those means receive equal model weights.

All 10,000 replicates preserve cue/source pairing and overlapping reset worlds. Per-model intervals use the same paired world draws. These intervals quantify world uncertainty for fixed checkpoints, not the training population.

A.1 Original-cohort heterogeneity and decomposition

The original-cohort clean-source interaction is +2.28 points [0.04, 4.55]. Both reference decompositions place its mean positive component in first-action selection: $2.28 = 2.38 - 0.10$ at Q_E, and $2.28 = 2.68 - 0.40$ at Q_T.

Both continuation-weighting intervals include zero. These signed accounting terms differ from D_w: their unresolved means do not establish identical continuations or absent internal memory use.

Changing from erased to clean source changes the interaction by -0.40 points $[-1.57, 0.68]$. Useful history in the source therefore coexists with an unresolved source effect on the interaction.

Model 2 contributes +9.11 points; the other four model intervals include zero. Omitting model 2 gives +0.57 $[-1.77, 2.94]$. Appendix A retains every model; the mean is not a training-population effect.

These models reach different states. The common-state receiver study in Appendix Q.3 controls selection for its own history assay; it does not remove state selection from this matrix's model comparison.

Secondary source and sample checks. Weighting the matrix by the original clean action sampled once per anchor gives +2.33 points [0.04, 4.65]. Using each posterior's sampled erased action gives +2.69 [0.59, 4.91]. Both are secondary to distribution weighting.

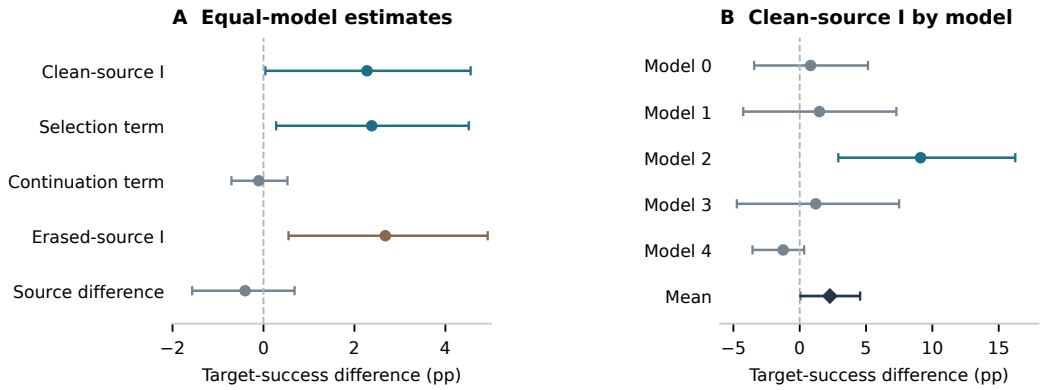

Figure 5: Original matrix decomposition and model heterogeneity. Left: equal-model components at the Q_E reference. Right: each model's clean-source interaction. Bars are conditional 95% reset-world intervals.

Selecting the native first-action sample instead of probability weighting changes free T success by +0.10 points [−0.19, 0.43] and E success by −0.08 [−0.54, 0.37]. These differences reflect finite sampling.

Changing the source from erased to clean changes native return by +0.037 [−0.012, 0.087] under T and +0.025 [−0.031, 0.079] under E. Neither interval resolves a native-return gain.

Model sensitivity. The leave-model-2-out calculation is exploratory and does not replace the specified five-model mean. Its clean-source interaction is +0.57 points [−1.77, 2.94]. The ordinary five-model t interval is [−2.65, 7.20].

Receiver 4 has free T/E target success near 48% on its own new panel. That local outcome is reported separately from the common-state visible control and cannot isolate a same-state memory limitation.

Raw-record verification. An independent reweighting of all raw matrices matches the stored per-anchor statistics to a maximum difference of 1.2×10^{-15}. Literal world resampling independently reproduces the three principal intervals.

All 428 extra sampled-action continuations match their corresponding matrix columns. This checks recorded statistics and selected execution paths; it does not repeat every simulator trajectory or establish cross-process invariance.

B Original-cohort post-review source and continuation analyses

Table 3: Original-cohort matrix-weighted target success. Sources weight each posterior's own continuations before averaging. Models have equal weight; intervals resample reset-world IDs jointly. The p_T fixed-source row was added after collection.

First-action rule	True (%)	Erased (%)	T minus E (pp, 95% CI)
Free, p_c	64.76	61.79	$+2.97\ [0.65, 5.28]$
Fixed, p_E	62.08	61.79	$+0.29\ [-0.89, 1.36]$
Fixed, p_T	64.76	64.17	$+0.59\ [-0.21, 1.43]$
Fixed, p_{clean}	67.36	66.67	$+0.69\ [-0.17, 1.56]$

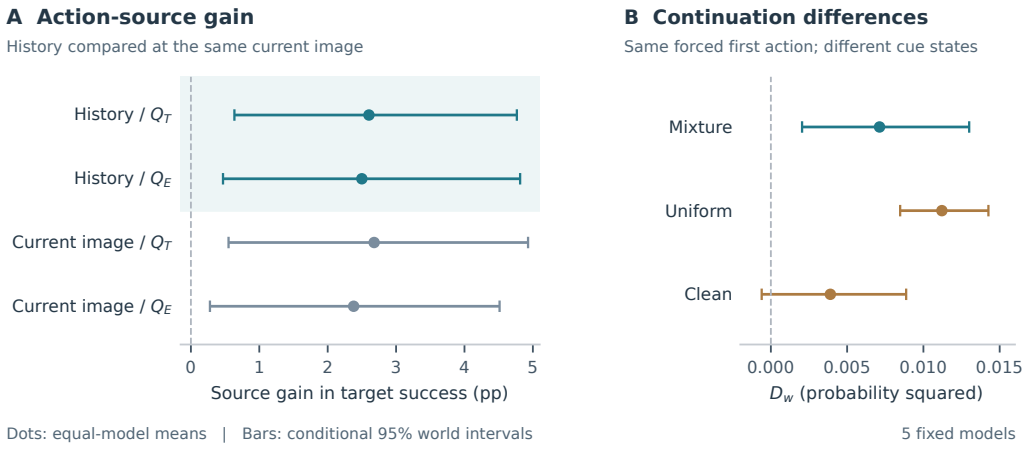

Figure 6: Original-cohort post-review analyses. Left: source gains under each continuation. Right: cross-repeat squared differences by action weight. Bars are conditional 95% world intervals for the original five checkpoints; panel units differ. Figure 2 reports the separate new-training cohort.

The analyses in this section reuse all 214 matrix anchors and 1,712 posterior draws. They were added after review. The calculation plan selected mixture weights before computing the new squared-difference summaries.

The same 10,000 joint resamples of 128 reset-world IDs used for the matrix apply here. Models remain equally weighted. All intervals condition on the fixed checkpoints. The original reanalyses retain the saved numerical execution; Section B.2 adds new source inference with saved tails.

B.1 Matching the action source's current observation

The p_T source already exists as the free true-cue distribution. Reusing it as common action weights requires no new rollout. It retains the erased prefix but observes the same true current image as p_{clean}.

Thus $(p_{\text{clean}} - p_T)^\top Q_c$ changes only the historical images supplied to the source. The source distributions weight each paired posterior's continuations; their carries do not replace the states generating Q_c.

Table 4: Current-image and history-source gains in target success (pp, 95% world intervals). Each row changes one source input while preserving the other. Continuation condition and action enumeration are shared.

Source contrast	Under Q_T	Under Q_E
Current image: $p_T - p_E$	$+2.68$ $[0.55, 4.93]$	$+2.38$ $[0.28, 4.51]$
History: $p_{\text{clean}} - p_T$	$+2.60$ $[0.64, 4.77]$	$+2.50$ $[0.47, 4.81]$

For source p_T, fixed T equals free T by construction. Its interaction therefore equals $(p_T - p_E)^\top Q_E$, the erased-reference selection term. This equality is not additional empirical evidence.

Model 1 contributes most to the history-source gain. Removing it leaves +1.49 $[-0.24, 3.26]$ points under Q_T and +1.74 $[-0.11, 3.70]$ under Q_E. Both intervals cross zero; the full five-model conditional result remains.

All five omission intervals for D_w remain positive. Its model sensitivity differs from that of the history-source gain. Neither exercise establishes generalization to an independently trained population.

Table 5: History-source target-success gains, $p_{\text{clean}} - p_T$, for each model (pp) and mixture-weighted D_w (probability squared), with 95% world intervals. Reached physical states can differ across models.

Model	Anchors	History, Q_T	History, Q_E	D_w
0	43	$+2.65$ $[-2.02, 7.45]$	$+2.50$ $[-2.20, 7.17]$	$+0.01025$ $[0.00154, 0.02092]$
1	52	$+7.06$ $[1.66, 13.30]$	$+5.55$ $[0.60, 11.36]$	$+0.00195$ $[-0.00697, 0.01107]$
2	42	$+3.85$ $[0.43, 7.87]$	$+5.03$ $[1.47, 9.03]$	$+0.00930$ $[-0.00000, 0.02425]$
3	25	-0.54 $[-4.20, 3.30]$	$+0.01$ $[-4.03, 4.75]$	$+0.00925$ $[-0.00489, 0.02924]$
4	52	-0.01 $[-1.05, 1.35]$	-0.59 $[-1.79, 0.75]$	$+0.00486$ $[0.00000, 0.01136]$

Table 6: Exploratory model omissions, with remaining models equally weighted and the same joint world bootstrap. History gains use pp; D_w uses probability squared. These conditional intervals do not sample new training runs.

Omitted	History, Q_T (pp)	History, Q_E (pp)	D_w
None	$+2.60$ $[0.64, 4.77]$	$+2.50$ $[0.47, 4.81]$	$+0.00712$ $[0.00205, 0.01300]$
0	$+2.59$ $[0.59, 4.80]$	$+2.50$ $[0.38, 4.88]$	$+0.00634$ $[0.00073, 0.01299]$
1	$+1.49$ $[-0.24, 3.26]$	$+1.74$ $[-0.11, 3.70]$	$+0.00842$ $[0.00257, 0.01535]$
2	$+2.29$ $[0.02, 4.88]$	$+1.87$ $[-0.50, 4.69]$	$+0.00658$ $[0.00100, 0.01333]$
3	$+3.39$ $[1.31, 5.71]$	$+3.12$ $[1.11, 5.31]$	$+0.00659$ $[0.00214, 0.01150]$
4	$+3.26$ $[0.74, 5.99]$	$+3.27$ $[0.65, 6.21]$	$+0.00769$ $[0.00163, 0.01485]$

Descriptive five-model t intervals for history gains are $[-1.23, 6.44]$ and $[-0.98, 5.98]$ points. Five fixed models, shared worlds and different reached states limit a training-population interpretation.

History-source native-return gains are +0.0187 $[-0.0127, 0.0513]$ under Q_T and +0.0147 $[-0.0199, 0.0485]$ under Q_E. The target-specific success result does not establish a native-control improvement.

B.2 Original-cohort historical content and spatial masking controls

We replay the original prefixes for every matrix anchor: 43, 52, 42, 25 and 52 for models 0–4, respectively, with eight draws each. All four sources see the same actual current image and recorded actions and rewards.

Only past images change. Clean and candidate-erased prefixes reproduce the original source conditions. A swapped prefix exchanges target/opposite crops where both are visible and unoccupied by the player or mobile entities.

The distant-mask source blackens the same number of 7×7 map tiles on each frame as candidate erasure. It selects the farthest available tiles from the candidate pair, with deterministic world-coordinate tie breaking.

It avoids the player, candidates, mobile entities and positions outside the map. Thus masking area and timing match, but visual content does not. Distant tiles are spatial controls; their irrelevance to all survival objectives is not assumed.

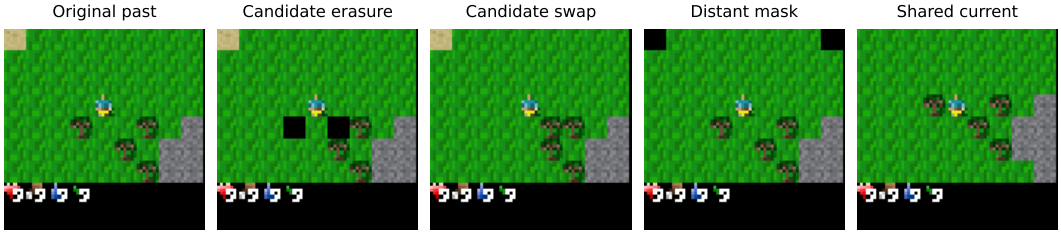

Figure 7: Actual source-control images from the first retained model-0 anchor. The last changed historical frame is shown; the current image at right is shared by all sources. Policy inference replays the entire recorded prefix.

All 214 anchors remain. Prefix lengths span 2–72 frames; every anchor has 2–63 changed swap frames. Candidate and distant masks cover 196–6,174 pixel locations summed across the past frames of an anchor.

Each new source distribution weights its paired draw's saved 17-action Q_T and Q_E. New clean and erased sources are extracted in the same process as the controls. Neither new tail rollouts nor new trained models enter this assay.

Table 7: Historical source controls: matrix-weighted target-success contrasts (pp, conditional 95% world intervals). Five checkpoints have equal weight; 10,000 joint resamples use the original 128-world indexing frame.

Source contrast	Q_T	Q_E
Clean − candidate erasure	+2.61 [0.64, 4.77]	+2.50 [0.47, 4.82]
Clean − candidate swap	+19.62 [14.77, 24.81]	+18.14 [13.55, 23.12]
Distant mask − candidate erasure	+3.48 [1.29, 5.83]	+3.53 [1.28, 6.04]
Clean − distant mask	−0.88 [-2.31, 0.37]	−1.03 [-2.47, 0.24]

The swap contrast supports sensitivity to past candidate content. The distant-mask contrast shows that erasing candidate locations has a different effect from erasing equal area at other locations on the same frames.

These findings narrow the generic-mask explanation, without identifying a pure semantic mechanism. Swapping introduces misleading content; its effect need not equal the benefit of clean over erased history.

Clean-minus-distant-mask intervals cross zero. This leaves that difference unresolved rather than establishing harmlessness or equivalence. The controls were specified before collection but remain exploratory on the original models.

Table 8: Target-success controls by model (pp, 95% world intervals). Swap is clean minus candidate swap; mask is distant mask minus candidate erasure. The two effects can have different model dependence.

Model	Swap, Q_T	Swap, Q_E	Mask, Q_T	Mask, Q_E
0	+17.69 [9.16, 27.76]	+16.95 [8.68, 26.60]	+5.16 [-0.92, 11.63]	+5.45 [-0.77, 12.27]
1	+36.52 [25.64, 47.42]	+34.45 [23.73, 45.24]	+6.58 [1.05, 12.75]	+5.18 [0.15, 10.70]
2	+19.95 [10.43, 30.30]	+20.24 [10.66, 30.66]	+4.51 [0.63, 9.31]	+5.54 [1.55, 10.24]
3	+14.26 [3.13, 27.21]	+9.88 [0.34, 21.57]	+0.67 [-3.39, 5.03]	+1.42 [-3.07, 6.49]
4	+9.68 [4.14, 16.23]	+9.15 [3.81, 15.47]	+0.49 [-1.21, 2.52]	+0.07 [-1.54, 1.89]

Table 9: The same source contrasts in native-return units, with conditional 95% world intervals. The positive clean-minus-swap contrast does not establish a clean-over-erasure native benefit, whose interval still crosses zero.

Source contrast	Q_T	Q_E
Clean − candidate erasure	+0.0186 [-0.0127, 0.0512]	+0.0147 [-0.0199, 0.0485]
Clean − candidate swap	+0.0691 [0.0174, 0.1238]	+0.0668 [0.0091, 0.1266]
Distant mask − candidate erasure	+0.0230 [-0.0097, 0.0572]	+0.0194 [-0.0139, 0.0525]
Clean − distant mask	−0.0044 [-0.0169, 0.0093]	−0.0047 [-0.0176, 0.0078]

Paired numerical sensitivity. The new clean-over-erasure gains are +2.6075 and +2.5033 points, versus archived +2.6032 and +2.4994. Paired changes are +0.00425 [−0.00567, 0.01433] and +0.00396 [−0.00578, 0.01387] points.

The maximum action-probability component drift is 0.03097; equal-model mean total variation is 0.000191 for clean and 0.000440 for erased history. Small endpoint drift coexists with nonidentical probabilities. Local reweighting of the downloaded raw sources reproduces cluster summaries to floating-point precision.

B.3 TWO REFERENCE DECOMPOSITIONS

Alongside Equation 14, the true-continuation reference gives

$$I_s = \mathbb{E}[(p_T - p_E)^\top Q_T + (p_E - s)^\top (Q_T - Q_E)]. \tag{4}$$

Table 10: Clean-source interaction components under two references (pp, 95% world intervals). Their sums agree; the individual components depend on the reference.

Reference	Selection	Continuation weighting
Q_E	$+2.38$ $[0.28, 4.51]$	-0.10 $[-0.70, 0.53]$
Q_T	$+2.68$ $[0.55, 4.93]$	-0.40 $[-1.57, 0.68]$

The observed mean positive component is first-action selection at either reference. This is an algebraic allocation of the contrast: recurrent state and physical action consequences both contribute to each continuation value.

B.4 WHY CROSS-REPEAT PRODUCTS REMOVE SQUARED ESTIMATION NOISE

Conditional on a posterior pair and action, write $d_r = \delta + \epsilon_r$. With zero-mean errors and independent repeats,

$$\begin{aligned} \mathbb{E}[d_1 d_2 \mid z_T, z_E, a] &= \delta^2, \\ \mathbb{E}[((d_1 + d_2)/2)^2 \mid z_T, z_E, a] &= \delta^2 + \tfrac{1}{4}(\operatorname{Var} \epsilon_1 + \operatorname{Var} \epsilon_2). \end{aligned} \tag{5}$$

The paired T/E outcomes within one repeat may be dependent. The required independence is between the two repeat pairs, conditional on the posterior pair, action, and fixed environment fork.

Weights depend only on the pre-tail action probabilities. The primary post-review choice is $(p_T + p_E)/2$; uniform and clean-source weights test sensitivity to the action weighting.

Table 11: Continuation differences in probability-squared units with 95% world intervals. Negative finite estimates and interval bounds are retained; no square roots or percentage-point conversions are applied.

Estimator and weight	Mean and 95% interval
Mixture, cross-repeat	$+0.00712$ $[0.00205, 0.01300]$
Mixture, naive square	$+0.02321$ $[0.01593, 0.03143]$
Uniform, cross-repeat	$+0.01121$ $[0.00848, 0.01426]$
Clean, cross-repeat	$+0.00390$ $[-0.00060, 0.00887]$

The same mixture weights give a signed difference of +0.44 points $[-0.45, 1.35]$. Its unresolved mean alongside positive D_w supports local continuation differences under the stated assumption.

The cross-product of each repeat's action-weighted mean is 0.00622 [0.00192, 0.01123]. Subtracting it from D_w estimates the within-posterior action-weighted variance under the same repeat-independence assumption.

This paired difference is 0.000908 $[-0.000452, 0.002932]$. It leaves within-posterior action variation unresolved. We cannot attribute the small signed mean mainly to cancellation between actions.

Table 6 reports every model omission. The clean-weight interval crosses zero, so a positive population squared difference is not resolved for every weight choice.

Random streams and calculation checks. The two tails use distinct folded environment keys and disjoint policy-counter intervals. The runtime derives policy seeds from the configured seed and counter. No recorded environment step key is reused across repeats.

These checks support the usual independent-pseudorandom-stream assumption; they do not prove independence mathematically. Two repeats leave appreciable noise, and numerical sensitivity requires the separate replay experiment.

Tests recover zero corrected expectation under an independent Bernoulli null and retain opposing action effects that cancel in the signed mean. The maximum dual-reference arithmetic residual is 2.67×10^{-15}.

C PRESPECIFIED CONFIRMATION ON NEW TRAINING SEEDS

The original cohort's post-review findings motivated a separate confirmation. Before collecting this cohort, we specified five new DreamerV3 training seeds 5–9, training-world seeds 105–109, size200m, replay ratio 512, and a budget of one million driver transitions. Every model is retained. After training reaches the requested budget, we select the latest complete saved checkpoint, without inspecting the history-use endpoints; actual saved counters appear in Table 13.

Interruption and recovery. All five training runs encountered host-memory exhaustion during checkpoint saving near 0.76–0.77 million transitions. Training resumed from each run's latest complete checkpoint with increased host RAM and the existing native resume mechanism. Agent parameters, optimizer and counter were restored with the available replay; simulator/driver and transient random/carry state restarted. This is not exact trajectory continuation. Total counter targets, seeds, architecture, replay ratio, endpoints and eligibility remained unchanged. All five resumed runs and subsequent evaluations completed successfully.

Fresh evaluation block. The unchanged first-eligible, nonfacing natural-tree rule is applied to 128 attempted reset worlds per model at new world seed 260930, with a 128-transition prefix limit. Eight posterior draws, two paired tail repeats, 17 forced actions, and 16 actions per continuation yield 116,960 rollouts. The 215 eligible model-world pairs span 95 distinct reset-world IDs. The trajectories execute 1,867,856 simulator steps; all 460 continuations shorter than 16 actions remain in the outcomes. Absent anchors are not replaced; models have equal weight despite unequal anchor counts.

Primary endpoints and uncertainty. The three specified endpoints are $(p_{\text{clean}} - p_T)^\top Q_T$, $(p_{\text{clean}} - p_T)^\top Q_E$, and D_w at $w = (p_T + p_E)/2$. Native return and other source contrasts are secondary. Draws are averaged within model-world pairs before pooling. Ten thousand bootstrap replicates jointly resample the shared 128-world frame and independently sample five training-model indices with replacement. World multiplicities preserve cross-model pairing and recompute each model's eligible denominator. All 10,000 replicates retain nonzero denominators.

We report conditional-world and joint model/world 95% intervals, plus 98.333% marginal intervals for a Bonferroni adjustment across the three primary endpoints. These bootstrap intervals approximate uncertainty with only five training units; they are not exact finite-sample coverage guarantees.

Table 12: Independent-cohort primary endpoints. Adjusted intervals are 98.333% marginal model/world bootstrap intervals. Units are pp for source gains and probability squared for D_w. The original cohort is not pooled into these estimates.

Endpoint	Mean	World 95%	Model/world 95%	Adjusted model/world
History / Q_T (pp)	5.13	[2.48, 7.80]	[1.28, 9.65]	[0.13, 10.85]
History / Q_E (pp)	5.36	[2.64, 8.12]	[1.41, 9.87]	[0.07, 11.08]
Mixture D_w	0.00891	[0.00078, 0.01951]	[−0.00645, 0.03218]	[−0.01044, 0.04079]

Both adjusted history-source intervals have lower bounds just above zero. For D_w, even the conditional-world adjusted interval [−0.00052, 0.02254] crosses zero; its unadjusted conditional interval alone does not establish a robust confirmation of all three primary endpoints.

Secondary outcomes and interpretation. Free true- and erased-cue target success are 65.06% and 61.29%. History-source gains have positive point estimates in every new model, but native-return intervals cross zero (Table 15). The larger source gains than in the original cohort do not

Table 13: All five new training seeds. Entries are per-model point estimates, not individual significance claims. Negative finite cross-repeat estimates are retained although the population D_w is nonnegative.

Seed	Saved steps	Anchors	History / Q_T (pp)	History / Q_E (pp)	D_w
5	998,730	31	4.36	6.70	0.03489
6	996,750	39	6.56	5.08	−0.01366
7	997,320	51	5.59	5.26	0.00606
8	999,460	65	1.23	1.76	0.01534
9	996,770	29	7.93	7.98	0.00192

Table 14: Descriptive leave-one-model-out means on the independent cohort. Positive means do not imply positive omission-specific uncertainty intervals.

Omitted seed	History / Q_T (pp)	History / Q_E (pp)	D_w
5	5.33	5.02	0.00242
6	4.78	5.43	0.01455
7	5.02	5.38	0.00962
8	6.11	6.26	0.00730
9	4.43	4.70	0.01066

measure improvement caused by training changes: both the training sample and evaluation-world block differ. There is no new full-task achievement evaluation for this cohort.

Table 15: Secondary native-return source gains in original return units. Neither continuation gives a resolved model/world benefit.

Native-return contrast	Mean	World 95%	Model/world 95%
History / Q_T	0.00522	$[-0.03089, 0.04375]$	$[-0.04420, 0.06455]$
History / Q_E	0.02096	$[-0.01085, 0.05650]$	$[-0.03327, 0.08419]$

Reproduction. The anonymous bundle includes compact per-draw action distributions, per-tail outcomes and a NumPy entrypoint for this cohort. Reweighting and repeating all 10,000 resamples reproduces all 84 reported metrics and their intervals to floating-point precision. This verifies analysis of the completed experiment; it is not an additional training or simulator replication.

D Posterior pairing and alternative execution measurements

This appendix makes the random pairing in Equations 2 and 3 explicit. The cross-draw analyses and bootstrap numerical check were added after reviewing the independent confirmation. They reuse the saved matrices and do not add trained models or simulator runs.

D.1 What is paired, and what is averaged

Conditional on anchor context X, the original protocol generates a tuple $(Z_{\text{clean}}, Z_T, Z_E) \sim \gamma_X$. It replays the recorded actions with matched policy random inputs. T and E use one common erased-prefix recurrent state, then observe different current images with the same posterior random input. Clean replays the unmodified history and observes the true current image. All three action distributions are read before tail execution.

Both cohorts record JAX 0.4.33. The RSSM uses 32 categorical latent variables with 64 categories each, a 0.01 uniform mixture, and float32 logits. The archived implementation calls `jax.random.categorical` on the last logit axis: Gumbel noise is added to logits and `argmax` selects a category in the learned output order (no category permutation). Thus matched random inputs specify this Gumbel-max coupling, not an arbitrary sampler with the same marginal distribution. The policy counter for a posterior update is $10^6 \cdot \text{world} + 10^4 \cdot \text{draw} + \text{decision index}$; tail

repeats add 1000 and $100\,\mathrm{repeat}$ before step offsets. Changing the sampler or cross-branch noise-coordinate alignment can change the coupling. A common consistent relabeling of both branches and their downstream mappings preserves the Gumbel-max joint law; relabeling the noise coordinates as well preserves each semantic sampled trajectory.

The continuation value $Q_c(X, Z_c, a)$ averages future policy and environment randomness from that particular state, with first action a forced. It does not average over Z_c. The simulator fork is identical across the compared conditions, but a forced action affects both subsequent physical state and the next recurrent update.

For G_c^γ, the source provides only action probabilities. In particular, $p_T^\top Q_T$ retains the native dependence between the true-cue action distribution and its own continuation state. Replacing this product with the product of posterior means changes the execution being evaluated; it does not correct bias in the original product.

D.2 A SQUARED PAIRED RESPONSE IS NOT A SQUARED MARGINAL RESPONSE

For fixed anchor-specific weights $\bar{w}_X(a)$, an alternative is

$$D_{\mathrm{marg}} = \mathbb{E}_X \left[\sum_a \bar{w}_X(a) \left\{ \mathbb{E}_{Z_T|X} Q_T(X, Z_T, a) - \mathbb{E}_{Z_E|X} Q_E(X, Z_E, a) \right\}^2 \right]. \tag{6}$$

This compares posterior-averaged values. In contrast, D_w^γ averages squared differences between paired sampled states and can use state-dependent weights. The quantities need not agree.

For example, take one action and $U \sim \mathrm{Bernoulli}(1/2)$ with $Q_T(U) = U$ and $Q_E(U) = 1 - U$. Both marginal value distributions are identical, and $D_{\mathrm{marg}} = 0$. Pairing the same U gives squared difference one; independent pairing gives expectation $1/2$; pairing equal values gives zero. Thus even identical branch marginals do not determine the paired square.

With two tail repeats, $d_r = \delta + \epsilon_r$ and conditionally zero-mean errors give

$$\mathbb{E}[d_1 d_2 \mid X, Z, a] = \delta^2 + \mathrm{Cov}(\epsilon_1, \epsilon_2 \mid X, Z, a). \tag{7}$$

The required independence is between repeat pairs conditional on the sampled states and action. T/E dependence within a repeat is allowed. Conditional repeat independence makes the covariance term zero; it does not eliminate finite-repeat variance or dependence on the chosen posterior coupling.

D.3 CROSS-DRAW EXECUTION USING THE SAVED MATRICES

Each anchor has $K = 8$ posterior draws. Write $\widehat{Q}_{c,j}$ for the mean of its two tail outcomes and $\Delta p_i = p_{\mathrm{clean},i} - p_{T,i}$. We compare the original same-draw source estimate with

$$\widehat{G}_c^{\mathrm{cross}} = \frac{1}{K(K-1)} \sum_{i \neq j} \Delta p_i^\top \widehat{Q}_{c,j}. \tag{8}$$

The source-history contrast remains paired internally, while its recipient uses a different draw from the same anchor. Its population target has an independently drawn source tuple and recipient state conditional on X. This deliberately changes the dependence present in the original native term. We retain the original native execution results unchanged.

For a separate continuation-pairing sensitivity, define $d_{ij,r}(a) = Y_{T,i,r}(a) - Y_{E,j,r}(a)$ and $w_{ij}(a) = (p_{T,i}(a) + p_{E,j}(a))/2$. The alternative estimator is

$$\widehat{D}^{\mathrm{cross}} = \frac{1}{K(K-1)} \sum_{i \neq j} \sum_a w_{ij}(a) d_{ij,1}(a) d_{ij,2}(a). \tag{9}$$

This changes the joint law of the T/E states, including their prefix samples, and the distribution of their mixture weights. It estimates a differently paired square, not Equation 6. Its magnitude cannot confirm or replace the original D_w endpoint.

All 56 ordered off-diagonal pairs reuse the same eight draws; they are not 56 independent observations. Environment keys are shared across draws within each tail repeat, while the two repeats use

Table 16: Exploratory posterior-pairing sensitivity in target success. Intervals are conditional-world 95% for the original cohort and joint model/world 95% for the confirmation cohort. Source gains use percentage points; squared differences use probability-squared units. Cross-draw pairing changes the execution definition, and its intervals have no multiplicity adjustment.

Cohort	Measurement	Within-draw pairing	Off-diagonal pairing
Original	G_T (pp)	2.60 [0.64, 4.77]	2.18 [0.41, 4.10]
Original	G_E (pp)	2.50 [0.47, 4.81]	2.42 [0.53, 4.55]
Original	D_w	0.0071 [0.0020, 0.0130]	0.0181 [0.0090, 0.0280]
Confirmation	G_T (pp)	5.13 [1.28, 9.65]	4.92 [1.02, 9.42]
Confirmation	G_E (pp)	5.36 [1.41, 9.87]	4.97 [1.02, 9.46]
Confirmation	D_w	0.0089 [-0.0064, 0.0322]	0.0283 [0.0137, 0.0501]

Table 17: Numerical check of joint model/world 98.333% intervals for the three original primary endpoints. The last column is the minimum and maximum lower endpoint across five 100,000-replicate streams, not a confidence interval for that endpoint.

Endpoint	Original 10,000	Pooled 500,000	Lower-endpoint range
G_T (pp)	[0.13, 10.85]	[0.31, 10.91]	[0.30, 0.34]
G_E (pp)	[0.07, 11.08]	[0.27, 11.05]	[0.25, 0.34]
D_w	[−0.01044, 0.04079]	[−0.00985, 0.04084]	[−0.00997, −0.00979]

distinct streams. We average the dependent pairs inside each anchor before resampling reset-world IDs jointly across models. The original cohort uses conditional world intervals; the confirmation cohort also resamples training models. Models retain equal weight and all eligibility misses and terminal outcomes are retained.

D.4 Results and their scope

Table 16 reports both cohorts. Source-history gains remain positive under the cross-draw execution rule. In the new cohort all five model means are positive for both source contrasts. The paired changes from the original execution are -0.22 [-0.82, 0.37] points under Q_T and -0.39 [-1.07, 0.34] under Q_E. These unresolved differences do not establish equivalence. In the original cohort, the Q_T gain decreases by 0.43 points, with a paired difference interval [-0.85, -0.02].

The squared estimate increases in both cohorts. In the new cohort the change is 0.01934 [0.00823, 0.03318]. This sensitivity is substantive: cross-draw pairing changes the posterior comparison and its mixture weights, so a positive alternative does not constitute replication of the original continuation result. Neither alternative measures D_{marg}.

D.5 Monte Carlo stability of the original bootstrap endpoints

The primary confirmation analysis remains the originally specified 10,000 resamples. We repeat the same world/model resampling procedure in five independent streams of 100,000 replicates each, using seeds 202609211–202609215. All 500,000 replicates have valid eligible denominators. We preserve the point estimates, model/world units, percentile method, and three-endpoint Bonferroni adjustment.

Both source-gain lower endpoints remain positive in all five streams; the original continuation-square interval continues to include zero. Increasing the resample count checks Monte Carlo endpoint stability. It does not supply additional trained models, guarantee small-sample coverage, or turn the exploratory pairing comparison into a prespecified experiment.

The new entrypoint reproduces the original paired means and 95% intervals to a maximum absolute difference of 1.39×10^{-17}. It also checks the off-diagonal sum against direct enumeration and reproduces the equal-marginals example above. The anonymous archive contains both cohorts' per-draw arrays, this analysis, and the stream-specific interval results.

Table 18: Known stochastic-state calibration, showing the shared-branch-uniform tail arrangement. The full 24-condition experiment also includes independent branch tails. Means are over 1,000 datasets; coverage is the fraction of ± 1.96 standard-error intervals covering the exact paired target. The equal-marginal rows show that a nonzero paired square can be induced by coupling even when $D_{\mathrm{marg}} = 0$.

Mechanism	Latent coupling	Exact D	Mean $\widehat{D}$	Wald coverage
Positive	Shared	.1800	.1809	94.9%
Positive	Independent	.2250	.2265	94.5%
Positive	Antithetic	.3600	.3607	93.8%
Equal marginal	Shared	.0000	.0000	100.0%
Equal marginal	Independent	.1800	.1808	95.1%
Equal marginal	Antithetic	.3600	.3604	94.4%
Asymmetric, weighted	Shared	.1475	.1471	94.2%
Asymmetric, weighted	Independent	.1719	.1714	94.2%
Asymmetric, weighted	Antithetic	.2450	.2443	94.2%

D.6 Known stochastic-state calibration

The production estimator was also tested on a separate known binary stochastic state model with exact targets. We varied the latent coupling (shared Gumbel coordinates, independent coordinates, and antithetic coordinates) and the within-repeat tail arrangement, including state-dependent policy weights. Across 24 prespecified conditions and 1,000 independent datasets per condition, the correct repeat-product estimator tracked the analytic paired targets; the largest absolute difference between a condition mean and its target was 0.001479.

This calibration validates the algebra, repeat-noise correction, and explicit dependence on coupling in the stated structural models. It does not establish that the Dreamer posterior has a uniquely correct coupling, reproduce its RSSM latent dynamics, or resolve the independent-cohort D_w interval. The raw datasets and reconstruction checks are in `stochastic_coupling_calibration/` in the reproduction archive, separately from the Dreamer matrices; the root entrypoint is `reproduce_stochastic_coupling.py`.

E Which actions account for the history-source gain?

This exploratory reanalysis uses all 215 independent-cohort anchors and all 116,960 saved continuation rollouts, including 557 terminal outcomes (460 before action 16). It adds no training or simulator observations. The source histories differ, while both source distributions see the same true current image; each recipient retains its own recurrent state.

Table 19 puts the local target gain in context. Target-tree success rises from roughly 65% to 70% under Q_T, whereas any-wood success is already about 89% and its change remains unresolved. The corresponding environment-reward gains also remain unresolved.

Table 19: Source values and their differences, averaged within draws, anchors, and then equally across the five models. Success levels are percentages and their gains are percentage points; native rewards use environment-reward units. Both success endpoints require survival through the horizon. Intervals are descriptive, unadjusted joint model/world 95% intervals.

Endpoint	Recipient	$p_T^\top Q_c$	$p_{\mathrm{clean}}^\top Q_c$	Gain [joint 95%]
Target success	Q_T	65.06	70.19	$+5.13\ [1.28, 9.65]$
Target success	Q_E	63.54	68.89	$+5.36\ [1.41, 9.87]$
Any-wood success	Q_T	88.60	88.88	$+0.28\ [-1.31, 1.68]$
Any-wood success	Q_E	88.44	89.08	$+0.64\ [-0.46, 2.01]$
Native reward	Q_T	2.2386	2.2438	$+0.0052\ [-0.0442, 0.0646]$
Native reward	Q_E	2.2189	2.2399	$+0.0210\ [-0.0333, 0.0842]$

Grouping actions by the decision geometry. We fix six exhaustive groups before inspecting their outcome contributions: the direction toward the target, the opposite direction, the other two directions, DO, NOOP, and the remaining eleven actions. Target and opposite directions are read from the recorded coordinates and checked against the recorded correct-turn action. No anchor is selected or ranked by its gain. Clean history shifts 11.76 percentage points toward the target direction. The mean within-draw total variation between sources is $+26.70$ $[20.58, 33.31]$ percentage points.

Attributing the gain without averaging away dependence. For each posterior draw, write $\Delta p = p_{\text{clean}} - p_T$ and $b_c = p_T^\top Q_c$. The contribution of action group A_g is

$$C_{g,c} = \mathbb{E}\left[\sum_{a \in A_g} \Delta p(a)\{Q_c(a) - b_c\}\right], \qquad \sum_g C_{g,c} = G_c. \tag{10}$$

The baseline is the value of the source from which probability is shifted. Centering makes each term compare its action with that source's expected outcome. A different action-independent baseline changes individual group contributions while leaving their sum unchanged. Thus the allocation is a descriptive explanation under the stated baseline, not a unique causal decomposition. Products are formed within each posterior draw; multiplying an average probability shift by an average action value would generally give a different answer.

Table 20: All geometry groups, with means and descriptive joint 95% intervals in percentage points. The target-direction contribution exceeds the total gain because other groups partly offset it. Tiny NOOP estimates round to zero. These exploratory group intervals have no multiplicity adjustment.

Action group	Probability shift	Target contribution Q_T	Target contribution Q_E
Toward target	$+11.76$ $[5.31, 18.73]$	$+6.09$ $[2.70, 10.33]$	$+6.50$ $[3.10, 10.68]$
Opposite direction	-6.64 $[-12.51, -1.37]$	$+0.10$ $[-0.99, 1.28]$	-0.10 $[-0.96, 0.37]$
Other directions	$+2.47$ $[-1.37, 6.69]$	-1.02 $[-3.54, 0.17]$	-0.93 $[-3.53, 0.29]$
Interact (DO)	-7.40 $[-14.98, -1.22]$	-0.12 $[-1.33, 0.65]$	-0.22 $[-1.51, 0.49]$
Do nothing	-0.00 $[-0.00, 0.00]$	$+0.00$ $[-0.00, 0.00]$	$+0.00$ $[-0.00, 0.00]$
Remaining actions	-0.19 $[-2.13, 1.52]$	$+0.07$ $[-0.04, 0.36]$	$+0.11$ $[-0.04, 0.43]$

The bootstrap reuses the original 128-world frame, 10,000 resamples, seed 20260914, and model-selection random stream, with equal model weights. The original target and native-return gains and intervals are reconstructed to floating-point precision. The additional endpoint and group analyses do not alter the prespecified confirmation or its three-endpoint adjustment. The archive supplies per-model estimates, all three endpoints' group contributions, conditional-world intervals, and unweighted group action values. Those group values provide context rather than a factorization of Equation 10. The pattern explains a change in local target choice; it does not establish improved complete-task return or distinguish information removal from harm caused by edited images.

F PROSPECTIVE CLEAN-RECIPIENT EXTENSION

The design was specified before collecting new continuations, after inspecting the original results. It uses the same five checkpoints and all 215 independent-cohort anchors, rather than new training or validation worlds. Both sources see the same true current image. Only their first-action distributions are supplied to the recipient, which retains its own clean-history state and sees unmasked future images. Relative to the primary experiment, both recipient history and future visibility change.

Nominated-tree and any-wood gains are this extension's two co-primary outcomes. Their joint Bonferroni 97.5% intervals are $[-2.988, 8.628]$ and $[-3.974, 1.751]$ points. Native return is secondary. Neither co-primary supports a positive finding. Training-seed gains for nominated-tree success were +3.073, -2.608, +7.585, +1.312, and +6.198 points for seeds 5–9, respectively; this spread is consistent with the limited precision of five models. The run retains 58,480 continuation trajectories, 933,981 simulator transitions, and 256 terminal tails, of which 217 terminate before action 16.

Each anchor retains eight posterior draws, all 17 forced first actions, two continuation repeats, and the original 16-action horizon and random inputs. Both source distributions are recomputed in the

Table 21: Source values under the clean-history, unmasked recipient. Success levels are percentages and their gains are percentage points; return uses environment-reward units. Both success endpoints require first-wood acquisition and survival through the horizon; nominated-tree success requires that tree.

Outcome	Edited source	Clean source	Gain [joint 95%]
Nominated-tree success	69.236	72.348	+3.112 [−1.937, 7.702]
Any-wood success	92.990	92.464	−0.526 [−3.261, 1.390]
Native return	2.566	2.604	+0.038 [−0.024, 0.108]

same execution as the clean recipient; they match the archived probabilities exactly. Outcomes and source probabilities are multiplied within each draw before averaging.

Ten thousand bootstrap replicates (seed 20260925) resample the original 128 reset-world IDs jointly across models, retaining eligibility masks, and independently resample the five equally weighted models. Model-t 95% intervals likewise span zero: $[-1.912, 8.136]$ for nominated-tree gain and $[-3.101, 2.050]$ for any-wood gain. Positive masked-condition results and unresolved clean-recipient results do not establish different gains across conditions. The original primary endpoints are unchanged. Compact raw outcomes, probabilities, the design, and reconstruction code are in the anonymous archive.

G Cross-process endpoint sensitivity and numerical divergence

G.1 Targeted diagnosis of three earlier discrepancies

Earlier shared-state and aligned-readout runs differ in 20 first actions for receiver 4, with maximum probability-component difference 0.10436. We selected three recorded cases for a targeted replay diagnosis.

Table 22: Three preselected draw-0 cases. The third is a tolerance-level control, not a bitwise-identical probability vector. Numeric action labels follow the environment's discrete action encoding.

Donor	World	Cue	Original max. probability gap	Original actions
0	194	Immediate true	0.104364	3 / 3
1	60	Immediate true	0.082176	2 / 4
0	4	Recent true	2.38×10^{-7}	2 / 2

Four fresh processes interleave shared/aligned/shared or the reverse order. Within each process, both entrypoints agree. Across processes, either call order can reproduce either of the two archived probability versions.

In these cases, images, executed-action feedback, rewards, random counters, incoming/outgoing RSSM features, and the actor's input features agree exactly. All 12,606,481 loaded actor parameters in 11 arrays also agree.

The first observed divergence is after identical actor features and weights, at the action-head probability calculation. Re-evaluating the saved features through the native carry reproduces each process's own version.

The runs share the GPU host, batch size, bfloat16 compute precision, actor configuration, and XLA flags. The observations narrow the location; they do not identify a particular numerical kernel or layer as the cause.

Earlier archived-diagonal substitution changes the common-state receiver gap only from -5.33 to -5.34 points. This is a limited sensitivity result, not proof that every endpoint is stable.

Table 23: Paired endpoint sensitivity on the fixed panel (pp). Range is the largest minus smallest of the three new-process means. Matrix and shared-state rows use subsets; they do not replace the full-study means.

Endpoint	Archive	Proc. 0	Proc. 1	Proc. 2	Range
B_{history}	9.5622	9.5622	9.5622	9.5622	0.0000
I_{clean}	0.5174	0.5085	0.5085	0.5143	0.0058
Clean minus erased, Q_T	2.3529	2.3388	2.3388	2.3439	0.0051
Clean minus erased, Q_E	2.9771	2.9655	2.9655	2.9718	0.0062
History source, Q_T	2.4597	2.4570	2.4570	2.4575	0.0005
History source, Q_E	2.7711	2.7677	2.7677	2.7682	0.0004
Receiver-4 gap	−7.2917	−7.5000	−7.5000	−7.5000	0.0000

The delayed study retains two differing archived action traces out of 1,128 comparisons; target/any-wood success and terminal outcomes agree. Its contrasts use contemporaneous branches, as do the new matrix comparisons.

G.2 A PRESPECIFIED PANEL FOR ENDPOINT SENSITIVITY

After review, we specified a bounded panel independently of the discrepant outcomes. All five checkpoints run in three fresh Python processes each, sequentially within one GPU allocation per checkpoint.

All runs use RTX 3090 GPUs on one node, JAX 0.4.33, bfloat16, batch size one, and the same XLA settings. Within each model, checkpoint, GPU allocation, random keys, policy counters, and program order remain fixed.

The natural-history assay retains all 94 archived anchors and four draws. For each matrix model and each shared-state donor, eligible world IDs are sorted and 12 anchors selected at equal-rank-bin midpoints.

For N eligible IDs, selected zero-based ranks are $\lfloor (2j+1)N/24 \rfloor$, $j = 0, \ldots, 11$. Selection does not use probabilities, success, effect sizes, or discrepancy magnitudes.

The matrix panel has 60 anchors, four posterior draws, two tail repeats, 17 actions and the original H16. The shared panel has 60 donor anchors, five receivers and eight draws; only the first transition is repeated.

Each process reconstructs the same archived prefix using recorded actions. Reset keys, positions, rewards and tasks are checked against the archive. Natural-policy anchor selection is not rerun.

All 15 processes complete. They produce 3,384 natural rollouts, 48,960 matrix rollouts and 36,000 shared-state first steps: 873,402 actual transitions. All 21 matrix early terminations are retained.

There are 360 additional sampled-action checks, adding 5,760 transitions. Every check matches its matrix column. An independent direct calculation of each process's mean interaction agrees to 5.21×10^{-18}.

G.3 PAIRED EFFECT CHANGES

Ten thousand reset-world resamples use the same multiplicities for every process, model and donor. Natural, matrix and shared assays use their original 32, 128 and 256 world-ID frames respectively.

The shared panel has one empty-donor resample, leaving 9,999 valid draws; the other assays retain all 10,000. Intervals describe conditional world uncertainty, not variation over execution environments or training populations.

B_{history} remains +9.5622 points in all processes. Its native-return contrast ranges from 0.183904 to 0.184404 return units. Unchanged success therefore does not imply every trajectory or secondary endpoint is identical.

The matrix interaction ranges from +0.5085 to +0.5143 points on this subset. Its largest change from the matched archive is -0.0089 points $[-0.0339, 0.0039]$. The full-study estimate of +2.28 is a different cohort mean.

Process 2 minus process 0 changes the subset interaction by +0.0058 points $[-0.00005, 0.01543]$. This is small relative to the subset's approximately $[-4.19, 5.55]$-point interval for the interaction itself.

The shared subset's receiver-4 gap is -7.50 points in all three processes, versus -7.2917 in the matched archive. The change is -0.2083 $[-0.6818, 0]$ points and retains the same direction.

The history-source gains vary by less than 0.0005 points across new processes. Their subset world intervals already include zero; the panel measures execution sensitivity, not a new full-sample significance test.

The same panel also estimates D_w. Its new-process range is only 3.44×10^{-8} in probability-squared units. The subset's own interval crosses zero, so this is numerical sensitivity, not an independent positive replication.

Table 24: Continuation-square sensitivity on the same 60-anchor subset, in probability-squared units. This unit differs from Table 23. All executions retain negative estimates and interval bounds.

Execution	D_w	95% world interval
Archive	0.001838295	$[-0.006397, 0.012554]$
Process 0	0.001838284	$[-0.006397, 0.012554]$
Process 1	0.001838284	$[-0.006397, 0.012554]$
Process 2	0.001838250	$[-0.006397, 0.012554]$

Table 25: Each model's clean-source interaction on its 12-anchor sensitivity subset (pp). Columns preserve the three process outcomes; no process or model is selected for its agreement with the archive.

Model	Archive	Proc. 0	Proc. 1	Proc. 2
0	-0.2975	-0.2975	-0.2975	-0.2975
1	-5.1492	-5.1492	-5.1492	-5.1492
2	10.5538	10.5538	10.5538	10.5826
3	-0.5050	-0.5496	-0.5496	-0.5496
4	-2.0150	-2.0150	-2.0150	-2.0150

G.4 Probability and trajectory differences persist

The first two new processes agree in the recorded comparisons. Process 2 differs from process 0 in two natural action trajectories, with unchanged target-success flags, and one of 16,320 matrix target-success flags.

The matrix maximum probability-component difference between these processes is 0.011862; 59 action sequences change. In the shared panel it is 0.050930, although all 12,000 sampled first actions agree.

Against the archive, the shared panel retains a maximum probability difference of 0.104364 and three changed first actions. The matrix has a maximum difference of 0.031984 and four to five changed target-success flags.

Archive comparisons also change program and compilation context and node. The same-program new-process comparisons isolate a narrower execution setting. Neither comparison identifies the numerical kernel responsible.

The measured endpoint changes do not materially alter interpretation on this panel. This is bounded sensitivity evidence, not bitwise invariance or proof that probability differences are small at every state.

H TRAINED CALIBRATION AND COMPARISON WITH MEMORY DIAGNOSTICS

The supporting cue interaction. Alongside G and D_w, we evaluate I_{clean}. First measure the T-minus-E outcome difference when each recipient chooses its own first action, then subtract that difference when both use the clean source's action probabilities. This free-versus-fixed cue interaction is distinct from comparing clean and erased source histories. The reset-every-frame policy has a 25-point interaction despite zero history gain and zero D_w; an interaction alone therefore does not establish memory use.

Table 26: Full trained-calibration measurements, means over ten seeds. Return is expected binary success; history gain uses Q_T. The interaction I_{clean} is retained here alongside the two main measurements.

Information path	Return (%)	I_{clean} (pp)	History (pp)	D_w exact
Reset each frame	74.992	24.99	0.00	0.00000
Reset after action	99.999	25.00	25.00	0.00000
Recurrent carry	99.998	12.50	12.51	0.06246

H.1 TASK, POLICY AND TRAINING

A balanced binary target determines the correct side. Four 8×8 grayscale frames encode a past target cue, a delay, the first decision, and the final decision. Stage markers and the executed route are visible.

The partially observed task always reveals the initial target and reveals the current cue in half the states. The fully observed twin reveals the target at every frame. Both share the same transitions and binary payoff.

The final observation includes the first action's physical route. Thus a policy can succeed by using history for its first action and following the route afterward, even if its recurrent carry is reset between the actions.

Each policy has a 64-to-32 tanh image encoder, a 32-unit GRU and a two-action softmax head. The three variants reset carry every frame, only before the final decision, or never within an episode. Parameters are learned in all cases.

For each variant and observability setting, seeds 0–9 initialize separate training runs. Each runs 1,200 Adam updates, batch 128, learning rate 0.003. A separate development seed 100 ran 600 updates before the reported training.

Training enumerates both first actions and both final actions to maximize expected reward. The objective mixes 75% policy-chosen routes with 25% uniform forced routes, so off-policy routes are represented during training.

A fixed 25% erased-history augmentation covers the source intervention in POMDP training. Standard POMDP evaluation retains the initial cue. The entropy coefficient is 0.005; action probabilities are learned without action labels.

The local CPU run fits all 60 models in 86.6 seconds using PyTorch 2.14.0 and one compute thread. The budget, ten seeds and reset mechanisms were written down before fitting the reported models. No seed is selected by its outcome.

H.2 COMMON EVALUATION AND EXACT CONTINUATIONS

Each model uses 256 shared nuisance-image worlds, each with four balanced target/current-visibility combinations. The same background noise and target states are used for every policy. These are 1,024 decision states per model.

Clean, T and E sources follow the manuscript's timing. Clean retains the past cue; T erases it but keeps the actual current image; E also erases the current cue. Each source weights the same T/E continuation matrices.

Both possible final actions are enumerated, giving exact expected binary success for each forced first action. This separates estimator sampling error from the trained policy's actual continuation difference.

Two independent pairs of uniform draws additionally sample binary returns from these conditional success probabilities. Within each pair, T/E share the draw. This is a Monte Carlo check against exact enumeration, not new Dreamer execution.

The recurrent mean is $D_w = 0.062459$, with seed range [0.062281, 0.062496]. The two-repeat estimate is 0.058795 [0.052252, 0.065437]; the naive square is 0.090338. Both reset controls give exactly zero for all three quantities.

The paired two-repeat minus exact difference is -0.003664 [-0.010207, 0.002978]. Intervals jointly resample the 256 nuisance worlds 10,000 times at seed 20260914, conditional on the ten trained policies.

H.3 POPGym Arcade comparators

We implement the discrete-action policy variants of pixel sensitivity and recall density in Wang et al. (2026), Equations (4π) and (5π). The observations, policies and evaluation states match our outcome measurements.

For pixel sensitivity, we sum squared image-gradient norms over the two action probabilities. Recall density instead sums absolute gradients and normalizes each frame's contribution by the total over the observed sequence.

First-action gradients hold the prefix fixed. Final-action gradients fix the executed route, then average over that policy's route distribution. They do not differentiate through the discrete sampled route.

Observability gap compares each architecture's separately trained MDP and POMDP expected returns. Memory bias subtracts the reset-every-frame MDP return from each memory-capable MDP return, pairing training seeds.

Table 27: Existing diagnostics on the same trained calibration. Recall columns are historical-frame shares in percent; pixel columns are squared gradient norms in scientific notation. Gap and bias use percentage points.

Information path	Recall, first	Recall, final	Pixel, first	Pixel, final	Gap (pp)	Bias (pp)
Reset each frame	0.00	0.00	0.00e+00	0.00e+00	25.007	0.0000
Reset after action	25.38	0.00	8.61e-11	0.00e+00	0.000	0.0003
Recurrent carry	38.47	33.24	1.71e-09	8.49e-10	0.000	-0.0000

The successful paths have near-zero observability gaps and memory biases. Historical recall is positive at the first decision for both, but only the recurrent path has historical sensitivity after the route is held fixed.

Thus the final gradient diagnostic also distinguishes these paths. Our additional measurements quantify outcome differences and conditional benefit; this comparison does not show that prior tools are incapable of diagnosing memory.

H.4 Historical sensitivity with redundant current information

An exploratory breakdown separates the already collected states by whether the current image contains the target. It uses all states in each stratum, with no refitting or selection of favorable nuisance worlds.

When the current cue identifies the target, historical recall remains positive but removing history barely changes expected action-source success. When the current cue is absent, the source contrast detects its substantial benefit.

Table 28: Current-visibility breakdown, averaged over all ten seeds. History gain uses Q_T; historical recall is the first-action gradient share. Current visibility was part of the task; this stratified analysis is exploratory.

Information path	Current cue	History gain (pp)	History recall (%)
Reset each frame	Visible	0.00000	0.00
Reset each frame	Absent	0.00000	0.00
Reset after action	Visible	0.00080	32.26
Reset after action	Absent	49.99844	18.50
Recurrent carry	Visible	0.00543	46.56
Recurrent carry	Absent	25.00621	30.38

A gradient share is relative sensitivity, not a probability of memory use or an outcome benefit. In these nearly saturated policies, the absolute gradient norms are small. They should not be read as equivalent to the finite source intervention.

H.5 Every training seed

Table 29: All ten POMDP seeds per information-path variant. Source gains and exact continuation differences use the learned probabilities; the final column adds the two-repeat Monte Carlo estimator. No failed or weak seed is excluded.

Seed	Information path	Return (%)	History (pp)	D_w exact	$\widehat{D}_w$
0	Reset each frame	74.9959	0.0000	0.000000	0.000000
0	Reset after action	99.9987	24.9994	0.000000	0.000000
0	Recurrent carry	99.9966	12.5084	0.062432	0.061527
1	Reset each frame	74.9964	0.0000	0.000000	0.000000
1	Reset after action	99.9989	24.9996	0.000000	0.000000
1	Recurrent carry	99.9986	12.5035	0.062477	0.061539
2	Reset each frame	74.9966	0.0000	0.000000	0.000000
2	Reset after action	99.9988	24.9997	0.000000	0.000000
2	Recurrent carry	99.9988	12.5002	0.062496	0.054688
3	Reset each frame	74.9868	0.0000	0.000000	0.000000
3	Reset after action	99.9988	24.9998	0.000000	0.000000
3	Recurrent carry	99.9987	12.5012	0.062491	0.054690
4	Reset each frame	74.9967	0.0000	0.000000	0.000000
4	Reset after action	99.9987	24.9995	0.000000	0.000000
4	Recurrent carry	99.9986	12.5005	0.062495	0.061525
5	Reset each frame	74.9923	0.0000	0.000000	0.000000
5	Reset after action	99.9987	24.9998	0.000000	0.000000
5	Recurrent carry	99.9988	12.5005	0.062494	0.061524
6	Reset each frame	74.9863	0.0000	0.000000	0.000000
6	Reset after action	99.9988	24.9997	0.000000	0.000000
6	Recurrent carry	99.9986	12.5073	0.062455	0.054694
7	Reset each frame	74.9950	0.0000	0.000000	0.000000
7	Reset after action	99.9988	24.9997	0.000000	0.000000
7	Recurrent carry	99.9986	12.5014	0.062490	0.061524
8	Reset each frame	74.9911	0.0000	0.000000	0.000000
8	Reset after action	99.9984	24.9994	0.000000	0.000000
8	Recurrent carry	99.9981	12.5313	0.062281	0.061552
9	Reset each frame	74.9826	0.0000	0.000000	0.000000
9	Reset after action	99.9987	24.9998	0.000000	0.000000
9	Recurrent carry	99.9986	12.5038	0.062482	0.054688

The resets are structural controls, not inferred mechanisms in Dreamer. The four-frame task is small and does not validate a learned-probe intervention, long-history retention, or independently trained Craftax agents.

I Exact two-decision counterexample

A two-decision task makes internal storage and its later use explicit. A balanced hidden target x is left or right. The agent sees its cue, chooses route a, then observes the physical route and chooses b.

Table 30: Exact enumeration in the two-decision calibration. Effect columns are percentage points. Storage reads the known internal variable after the true cue; it is not a fitted probe. B_{history} fixes the route, whereas I permits cue-dependent route choices in its free arms.

Known policy	Storage (%)	B_{history}	I_{clean}	I_E
Memoryless	50	0	50	50
Store and use	100	25	50	25
Store and ignore	100	0	50	50

An unobserved fair coin chooses whether reward tests $a = x$ or $b = x$. All three policies choose $a = x$ with a true cue and choose uniformly after erasure. They differ only in storage and the second decision.

The memoryless policy has constant internal state and sets $b = a$. The store-and-use policy records x and sets $b = x$ when available. The store-and-ignore policy records x but sets $b = a$.

After erasure, the latter two policies also follow the visible route. For the matched-prefix history measurement, we force the same balanced route independently of x, leaving the second observation identical across histories.

All policies have free T/E success of 100/50%. A clean source always selects the correct route, giving fixed T/E success of 100/100%. Consequently, all three clean-source interactions are 50 points.

With an erased source, fixed T/E success is 50/50% for the memoryless and store-and-ignore policies, and 75/50% for store-and-use. The last interaction is 25 points, decomposing as $50 - 25$ at the Q_E reference.

Thus a nonzero interaction alone cannot identify internal memory. In this construction, storage plus the matched-prefix history effect distinguishes the three mechanisms; changing the source reveals a difference hidden by I_{clean}.

The calibration validates properties of the estimators on explicit policies. It does not assign these mechanisms to Dreamer or validate a learned probe pipeline. Dreamer inference requires the controlled behavioral measurements below.

J Baseline evaluation and metric definitions

J.1 Evaluation and metric definitions

The current baseline uses the public DreamerV3 implementation with a Craftax-Classic pixel wrapper. Native 63×63 images are padded to 64×64 without resizing the tile grid. The environment has 17 discrete actions and 22 achievements. The evaluator loads a frozen checkpoint and scores 100 complete episodes. For Craftax, world generation is a pure function of evaluation seed and episode index. Evaluation samples from the learned stochastic policy; it is not greedy.

For achievement indicator x_{ej} in episode e, we report three distinct quantities:

$$P = \frac{100}{22N} \sum_{e=1}^{N} \sum_{j=1}^{22} \mathbf{1}[x_{ej} > 0], \tag{11}$$

$$S = \exp\left(\frac{1}{22} \sum_{j=1}^{22} \log\left(1 + \frac{100}{N} \sum_{e} \mathbf{1}[x_{ej} > 0] \right) \right) - 1, \tag{12}$$

$$R_{\text{norm}} = \frac{100}{22N} \sum_{e} R_e. \tag{13}$$

P is mean achievement percentage; S is the geometric achievement score; R_{norm} includes the environment's health shaping. Averaging a geometric score computed separately within episodes would give a different statistic.

Table 31: Original-cohort Craftax-Classic baseline at replay ratio 512, with 100 evaluation episodes per training seed. World seed denotes the training-world seed; driver counts include resets. All three metrics use the same saved episodes.

Agent seed	World seed	Driver count	P (%)	S	R_{norm} (%)
0	100	996,810	48.27	15.54	44.18
1	101	998,640	53.00	19.61	48.91
2	102	999,130	57.68	25.15	53.59
3	103	994,950	45.05	14.06	40.95
4	104	997,610	48.45	14.10	44.36

Dedieu et al. report a Craftax-Classic DreamerV3 reference of 47.18 ± 3.88 in their Reward (%) column, with uncertainty defined as standard error over 10 seeds (Dedieu et al., 2025). Their normalized reward is compared with our R_{norm} column, rather than achievement percentage P; remaining health and evaluation-window conventions still limit direct implementation comparison. The implied seed standard deviation is about 12.27. The reference is a comparison point, not a known population mean, and agreement of a point estimate does not establish statistical equivalence.

The configuration is named `size200m`; the live implementation reports 165,965,072 optimized parameters. This is not a parameter-matched replication of the published 201M implementation. Exploratory checkpoints and their training-provenance limitations are detailed in Appendix Q.1.

J.2 Original-cohort baseline evaluation

The original five agents use seeds 0–4, separate training-world seeds 100–104, replay ratio 512, and a requested budget of one million driver transitions. Loaded checkpoint counters range from 994,950 to 999,130 (Table 31). Each checkpoint was evaluated for 100 episodes with evaluation seed 12345 and policy seed 0, retaining its restored policy action counter.

The equally weighted mean achievement percentage is **50.49%**, with seed SD 4.92, SEM 2.20, and a seed-bootstrap 95% interval of [47.02%, 54.86%]. Mean S is 17.69 (seed SD 4.74; SEM 2.12), and mean R_{norm} is 46.40% (seed SD 4.92; SEM 2.20). The 4.09-point difference between P and R_{norm} reflects health shaping in these episodes.

K Natural history protocol and detailed outcomes

Retaining earlier candidate-tile views increases target first-wood success from 71.14% to 80.70% with the same masked current observation, a mean gain of 9.56 percentage points across the five original-cohort agents (seeds 0–4). The difference is positive for every seed (Table 32); unmasked observations give 81.28% success.

Natural anchors and executed outcomes. Each agent follows its own unedited policy from the same 32 reset-world seeds. Before each next action, we select the first state with exactly one adjacent tree, no previously collected wood, and at least two earlier observations exposing the target and opposite tiles without player or entity overlays. The search uses past and current state only, with a 128-transition limit. It yields 94 eligible seed–world pairs among 160 examined: 65 already collected their first wood before eligibility, and one terminated during the prefix. The selected decision indices range from 2 to 61, with median 2. These are policy-specific anchors on shared world seeds, rather than 94 independent worlds.

Paired history intervention. We replay the recorded prefix with the actual executed actions, either retaining the original images or masking the fixed target and opposite coordinates from reset. At the anchor, both branches receive the same masked current image, previous action, reward, and simulator state. A visible branch retains the clean prefix and current view. Four policy draws share random counters and simulator keys across conditions; subsequent native actions respond freely for 16 steps. No simulator fields are edited. The mask follows fixed world coordinates and leaves a player-occupied tile visible. We record whether the first wood came from the target, whether it came

Table 32: Natural first-wood success (%) within 16 executed actions. Δ is seen–masked minus erased. Four draws are averaged within each eligible world, then worlds within a seed; the final row weights seeds equally.

Seed	Anchors	Visible	Seen–masked	Erased	Δ (pp)
0	22	92.05	89.77	82.95	6.82
1	22	84.09	85.23	73.86	11.36
2	20	100.00	95.00	87.50	7.50
3	11	75.00	79.55	61.36	18.18
4	19	55.26	53.95	50.00	3.95
Mean	—	81.28	80.70	71.14	9.56

from any tree, and cumulative native return. There are 376 trials per condition, 1,128 total, with no terminal evaluation branches.

Target choice and broader performance. The paired target effect has seed SD 5.50 percentage points and a seed-bootstrap 95% interval of [5.81, 14.09] points. These intervals condition on the recorded eligible anchors and are coarse with five training seeds. Any-wood success is 96.46%, 94.50%, and 92.30% for visible, seen–masked, and erased conditions. Its paired gain is 2.20 points, positive in two seeds, zero in two, and negative in one; its bootstrap interval is $[-1.12, 6.82]$ points. Mean native returns are 2.871, 2.490, and 2.305, respectively. The seen–masked return gain is 0.184, positive in four seeds. Its bootstrap interval is [0.029, 0.343], whereas a small-sample t interval is $[-0.066, 0.435]$, so a broad return benefit remains uncertain. The strongest observed effect concerns which tree supplies the first wood, conditional on these natural approaches.

Waiting changes visible competence. A follow-up on these same 94 anchors adds 0, 2, 8, or 16 real NOOP transitions. History effects remain positive but nonmonotone, while visible target success falls from 81.28% to 43.59%. Replaying the full visible history at delay 16 reduces target success by 10.25 points relative to seen–masked history, with negative effects in all five agents. The 8,832-rollout diagnostic therefore does not identify a pure forgetting curve or a successful restoration mechanism. Its full methods, control results, and uncertainty are retained in Appendix Q.9. The next assays use natural decision states without adding a NOOP waiting period.

L One-transition history replication

L.1 A cue one natural transition before the decision

The nonfacing natural anchor must expose both candidates at the preceding observation, with unchanged map classes through the current decision. The inherited two-exposure rule puts this cue after reset. Recent true/mismatched histories reveal only $t-1$; the latter swaps the candidate crops. Both receive a masked current image. Erasure masks both candidates throughout, while the immediate control shares that erased history and reveals the actual image once at t. All later views are masked. The source action and current physical state are fixed, and the primary recent-minus-erased endpoint is the first executed target-directed turn. Timing and viewpoint both differ in the recent/immediate comparison (Figure 4A).

The established development seed 0 yields 78 eligible approaches from 256 worlds. Recent-minus-erased first turn improves +6.57 points [3.37, 10.26], with an immediate-cue effect of +7.21 [3.04, 11.86]. Full protocol details and secondary seed-0 outcomes are in Appendix Q.2. We then replicated the unchanged protocol on the other four checkpoints.

Replication across four additional agents. Each new agent uses the specified 256-world block at reset seed 260912, eight draws, and 16 free actions. The executed first-turn endpoint remains primary; policy probability is supporting evidence. Seeds 1–4 contribute 92, 65, 48, and 101 eligible approaches, respectively: 306 seed–world pairs from 1,024 attempts, 9,792 rollouts and 156,672 native steps, with no rollout termination. The other attempts comprise 715 first-wood collections before eligibility, one prefix-limit exclusion, and two deaths during the source prefix. These pre-

Table 33: Unchanged one-transition replication on new agents 1–4. Gains are recent true minus erased. Each agent's interval resamples its paired eligible worlds; the last row weights agents equally and sums their anchor counts. The prior seed-0 block is excluded from this primary aggregate.

Seed	Anchors	First turn gain points [world 95% CI]	Policy prob. gain (points)	Target success gain (points)
1	92	8.02 [4.21, 12.09]	7.89	4.21
2	65	5.38 [2.50, 8.46]	6.41	1.35
3	48	3.91 [1.04, 7.29]	3.00	0.52
4	101	-0.37 [-1.61, 0.74]	-0.35	0.74
New four	306	4.23	4.24	1.71

intervention exclusions are retained without replacement. Checks of raw actions, acquisition events, rewards, and saved image panels pass.

Recent true raises equal-seed first correct turn from 33.24% to 37.47%: **+4.23 points** (seed SD 3.51; SEM 1.75). The three positive agent effects have conditional world intervals above zero (Table 33). Seed 4 gives -0.37 points $[-1.61, 0.74]$, without evidence of benefit or harm. Policy probability changes by +4.24 points with the same three-positive, one-negative pattern. The ordinary four-seed t sensitivity interval for the first-turn mean is $[-1.35, 9.82]$ points; independent-agent population precision remains limited despite the resolved within-agent effects.

The same-cohort immediate control gives +8.83, +16.54, +5.47, and 0.00 first-turn points in seeds 1–4, respectively. Thus the seed-4 history exception also lacks a resolved immediate benefit over erasure. Recent true nevertheless exceeds mismatch for first turn and target success in every agent, with conditional world intervals above zero; the equal-seed gains are +7.52 and +7.57 points. Content sensitivity and benefit over erasure remain distinct.

Recent-minus-erased target success changes by +1.71 points (four positive point estimates; t sensitivity interval $[-1.01, 4.42]$); only seed 1 has a conditional world interval above zero. Target acquisition penalty decreases 0.329 steps, while any-wood success and native return change by +0.42 points and +0.0114. These secondary effects do not establish a broad control improvement.

Recent minus immediate first turn is -3.48 points on average. Seed 2 has a resolved conditional difference of -11.15 points $[-16.54, -5.96]$; the other agents' intervals include zero. Timing and viewpoint both change, so this is not an isolated forgetting estimate. The overlapping all-five summary adds the prior seed-0 block: 384 eligible pairs and 12,288 rollouts, with +4.70 first-turn points (four positive agents; seed SD 3.21, SEM 1.44). Its ordinary t sensitivity interval is [0.71, 8.69], and it remains a secondary summary. The new results extend conditional one-transition history use to three additional agents while retaining heterogeneity and limited population precision.

M ORIGINAL SOURCE-ACTION REPLICATION: DETAILED CONTRASTS

The sampled-source study motivated the subsequent continuation matrix. Table 34 retains its absolute outcomes. These cohorts and once-per-anchor actions differ from the matrix's source weighting.

Table 34: Absolute 16-action target success in the original four-agent replication, with equal agent weights. The clean source observes unedited past and current images; its action is sampled once per anchor.

Original first-action rule	True (%)	Erased (%)	Mismatched (%)
Fixed clean-source sample	72.54	72.27	70.11
Free branch sample	71.60	66.42	52.44
Fixed minus free (pp)	+0.94	+5.85	+17.67

We specified a fresh 256-world development block at the established seed-0 checkpoint, using non-facing natural anchors and six cue/execution arms at the cue-frame fork sketched in Figure 4B. The source policy's original action is sampled and recorded after eligibility, without inspecting its

Table 35: Primary true-minus-erased target-success effects at 16 actions. The interaction is free minus fixed, paired across all four arms within each world–draw. Intervals resample each agent's eligible worlds. The final row weights new agents equally and excludes prior seed 0.

Seed	Anchors	Fixed cue effect (pp)	Free cue effect (pp)	Interaction pp [world 95% CI]
1	87	-1.01	2.59	3.59 [0.86, 7.04]
2	74	0.68	9.29	8.61 [4.90, 13.01]
3	53	1.42	9.20	7.78 [3.07, 13.21]
4	107	0.00	-0.35	-0.35 [-3.04, 2.45]
New four	321	0.27	5.18	4.91

consequences. Fixed arms execute that action only at the cue frame; free arms execute their own sample. All arms share the erased incoming history, physical fork state, past actions and rewards, policy counters and 16 transition keys. Every later action executes freely, with candidate masking. No branches are selected by subsequent success.

The primary endpoint is the paired true-minus-erased target-success interaction I, with the true-minus-mismatched interaction secondary. Acquisition windows at 2/4/8/16 steps and all-trial acquisition penalties were specified before execution. The fixed-arm executed first-turn effect is zero by construction, so its contrast with free first turns cannot itself establish a downstream methodological finding.

The seed-0 development block yields 77 eligible approaches and an unresolved primary interaction of -0.97 points $[-4.38, 2.44]$. Its first-turn cue effect is positive, and opposite trajectory examples show both attenuation and amplification under fixed actions (Appendices Q.5 and Q.6). We retained this result and specified unchanged replication on all other four agents before collecting their outcomes, because a one-agent null does not settle the cross-agent question.

Unchanged replication across four new agents. Seeds 1–4 each complete the same 256-world block (reset seed 260913), eight draws and six arms. They yield 87, 74, 53 and 107 anchors: 321 eligible seed–world pairs, 15,408 rollouts and 246,528 native steps, with no rollout deaths. The other 703 attempts comprise 700 first-wood collections before eligibility, two source-prefix deaths and one prefix limit. All 7,704 within-cue fixed/free first samples and probabilities agree, and all 5,068 same-executed-action pairs have identical complete trajectories.

The equal-new-agent cue effect is +0.27 target-success points under fixed execution and +5.18 under free execution. Their primary interaction is **+4.91 points** (seed SD 4.14; SEM 2.07). Seeds 1–3 have positive conditional world intervals, while seed 4 is unresolved (Table 35). The ordinary four-seed t sensitivity interval is $[-1.68, 11.50]$ points. Thus attenuation under fixed actions replicates conditionally in three additional agents, without a precise training-population estimate or a universal effect. In seed 1, the interaction is resolved even though each mode's individual true-minus-erased interval includes zero; comparing those individual significance decisions would miss it.

The secondary true-minus-mismatched target-success interactions are 20.26, 20.78, 18.16 and 7.71 points, with positive conditional world intervals in all four agents. Their equal-seed mean is **16.73 points** (ordinary t sensitivity interval [6.99, 26.46]). Fixing the source action therefore attenuates this content contrast more consistently than the benefit-over-erasure contrast. This cannot establish universal utility of the true cue, or replace the primary comparison.

Primary-control acquisition interactions are +7.70, +5.88, +5.53 and +4.91 points at 2/4/8/16 actions; all four ordinary t intervals include zero. The target acquisition-penalty interaction is -0.850 steps, while native return and any-wood success interactions are +0.0230 and +1.39 points, also imprecise across agents. Absolute true-cue target success is 72.54% under fixed execution and 71.60% under free execution: a larger cue contrast does not itself imply a better controller. Appendix Q.7 reports all six condition means.

The overlapping all-five summary includes 398 eligible pairs and 19,104 rollouts. Its primary interaction is +3.73 points (three positive and two unresolved agents; ordinary t interval $[-1.79, 9.25]$). The prior null, seed-4 exception, and opposing individual trajectories remain part of the evidence.

The result supports reporting the action constraint and cue control alongside the measured effect, rather than treating fixed replay as an error or free execution as an automatically superior measurement.

N MEDIATION CONNECTIONS AND SUPPORTING CONTRASTS

N.1 RELATION TO DIRECT AND INDIRECT EFFECTS

Direct, indirect and interaction effects are established mediation quantities (VanderWeele, 2014; VanderWeele et al., 2014). For the cue comparison, let treatment A choose T or E, $M(A)$ be the first action sampled from p_A, and $Y(A, m)$ the endpoint obtained by retaining Z_A and forcing action m. The simulator noise protocol induces the state-tuple law γ_X; augment it with first-action sampling noise independent of tail noise conditional on the state tuple. Then the natural-direct-effect form is $\mathbb{E}[Y(T, M(E)) - Y(E, M(E))] = \mathbb{E}[p_E^\top (Q_T - Q_E)]$; the total-indirect-effect form is $\mathbb{E}[Y(T, M(T)) - Y(T, M(E))] = \mathbb{E}[(p_T - p_E)^\top Q_T]$. Their sum is the total cue contrast. Using p_T for the direct contrast instead gives the complementary pure-indirect-effect form at Q_E.

These are defined simulator interventions. Since Z_A affects both the mediator and subsequent behavior, the usual observational mediation formula does not identify them without additional assumptions. The executable branch and specified noise coupling supply additional structure; numerical equality of formulas alone does not identify a unique natural effect. For history as treatment and a true current cue at the decision in both history conditions, followed by the same masked-continuation protocol, G_T has the pure-natural-indirect-effect form. G_E additionally fixes an erased-current recipient and therefore needs the expanded hybrid intervention specification. The cross-draw source comparison instead independently samples source and recipient. The squared response D_w is a second-moment diagnostic, not a natural indirect effect.

N.2 REFERENCE-DEPENDENT CUE ACCOUNTING

Define the cue gain $\Delta_m = \mathbb{E}_\gamma[V_T^m - V_E^m]$ and interaction $I_s = \Delta_{\text{free}} - \Delta_{\text{fixed(s)}}$. Expanding the products gives

$$I_s = \mathbb{E}_\gamma \left[\underbrace{(p_T - p_E)^\top Q_E}_{\text{first-action selection}} + \underbrace{(p_T - s)^\top (Q_T - Q_E)}_{\text{continuation weighting}} \right]. \tag{14}$$

The first term evaluates the action-distribution shift using erased-cue continuations. The second weights same-action continuation differences by the remaining distribution difference. These terms depend on the chosen reference.

This is a reference-dependent accounting identity. The selection term also includes subsequent recurrent behavior: a forced action changes the physical state and enters the next RSSM update. It is not a pure external-memory path.

N.3 HISTORY UNDER MATCHED PREFIX ACTIONS

Replay an eligible trajectory with informative (H^+) or erased (H^-) candidate-tile history. Executed prefix actions, current simulator state, previous reward, and masked current image are shared.

For an executed endpoint G, the history effect is

$$B_{\text{history}} = \mathbb{E}[G(H^+) - G(H^-) \mid \text{prefix actions and eligibility}]. \tag{15}$$

Subsequent actions respond freely. This effect measures the contribution of selected past images under the recorded actions. Those actions can themselves carry information, so erasure does not remove every source of target evidence.

O SAVED-MATRIX SENSITIVITY AND RECORDED CONTEXT

These exploratory analyses were specified after the September 23 review and reuse the two saved matrix cohorts. They retain every eligible anchor and terminal outcome. The three primary es-

Table 36: Exploratory fixed-weight squared responses in probability-squared units; joint model/world 95% intervals without multiplicity adjustment. Weights use leave-two draws for the marginal response and leave-one for the paired response. Conditional-world intervals for D_{marg} are $[0.00020, 0.00630]$ and $[0.00090, 0.01314]$, respectively.

Cohort	D_{marg}	$D^{\gamma}(\bar{w}_X)$
1	$0.00287\ [-0.00094, 0.00823]$	$0.00637\ [-0.00001, 0.01487]$
2	$0.00587\ [-0.00056, 0.02042]$	$0.01032\ [-0.00426, 0.03369]$

timates, their original bootstrap intervals, and the confirmation's multiplicity adjustment are unchanged.

A posterior-marginal response with independent weights. Let $w_i = (p_{T,i} + p_{E,i})/2$ and let $\bar{w}_X = \mathbb{E}[w_i \mid X]$. For $K = 8$ draws, define $\bar{w}_{-ik} = (K-2)^{-1} \sum_{j \notin \{i,k\}} w_j$ and $\bar{w}_{-i} = (K-1)^{-1} \sum_{j \neq i} w_j$. An anchor-level estimator of Equation 6 is

$$\widehat{D}_{\mathrm{marg},X} = \frac{1}{K(K-1)} \sum_{i \neq k} \sum_{a} \bar{w}_{-ik}(a) d_{i,1}(a) d_{k,2}(a). \tag{16}$$

Averaging $\sum_a \bar{w}_{-i}(a) d_{i,1}(a) d_{i,2}(a)$ over i estimates the paired square $D^{\gamma}(\bar{w}_X)$ with the same population weights. These statements require independent posterior tuples conditional on X and independent complete tail-repeat blocks conditional on X and all tuples. Environment randomness may be shared across draws within a repeat. The 56 ordered pairs are averaged inside an anchor, not treated as 56 independent observations.

For these fixed weights, Jensen's inequality gives $D^{\gamma}(\bar{w}_X) \geq D_{\mathrm{marg}}(\bar{w}_X)$; their difference is the weighted posterior variance of $Q_T - Q_E$. This inequality does not give a bound using the original state-dependent D_w, and finite unbiased estimates need not preserve nonnegativity or ordering. Both new joint model/world intervals for D_{marg} include zero (Table 36).

Model-level sensitivity and secondary endpoints. For a like-for-like comparison with the confirmation, joint model/world resampling of the original cohort gives 95% intervals of $[-0.66, 6.42]$ and $[-0.72, 6.19]$ percentage points for G_T and G_E, and $[0.00033, 0.01578]$ for the original state-dependent D_w. These exploratory intervals reuse the saved matrices; they neither replace the original conditional-world analysis nor turn it into a prespecified test.

A Student-t sensitivity analysis over the five model means uses four degrees of freedom. Cohort 1's 95% intervals for G_T/G_E are $[-1.23, 6.44]/[-0.98, 5.98]$ percentage points; cohort 2's are $[1.97, 8.29]/[2.46, 8.25]$. Cohort 2's 98.333% intervals are $[0.62, 9.64]/[1.23, 9.49]$. This small-sample normal-model approximation conditions on the finite evaluation frame, is not inherently more conservative than the joint bootstrap, and does not replace the specified primary analysis. The completed cohort-2 analysis also reports the secondary current-source gains at Q_T/Q_E as $2.08/2.25$ points, with joint 95% intervals $[-0.64, 5.76]/[-0.51, 6.13]$, and $I_{\mathrm{clean}} = 2.47\ [-0.44, 6.46]$ points. All three conditional-world intervals are positive, but all three joint intervals include zero.

Prefix context, sparse estimates, and terminal outcomes. Cohorts 1 and 2 contain 214 and 215 anchors. Their prefix-action medians are both 3, with ranges 2–72 and 2–67; 39.3% and 43.3% occur after exactly two actions, and 74.8% and 65.6% after at most five. These are prefix lengths, not retention delays. The last *eligible visible exposure* is immediately before the decision at 213/214 and 215/215 anchors; the remaining cohort-1 anchor has 19 intervening observations. This exposure criterion is distinct from the last frame affected by the masking operation. The current target is visible by eligibility. Full per-model histograms and descriptive prefix strata are included in the reproduction output.

Using an explicitly post-hoc descriptive cutoff $|\widehat{D}_{w,X}| > 10^{-4}$, cohort 2 has 5, 5, 5, 8, and 3 contributing anchors by model. Exact nonzero counts are 22, 21, 30, 42, and 21 because many contributions are tiny; neither count estimates the number of nonzero population effects. In seed 5, world 126 contributes 0.01990 to the model mean of 0.03489.

There are 23/116,416 and 557/116,960 terminal matrix rollouts; 21 and 460 terminate before action 16. Cohort 2 includes 171 rollouts that collect the target but then fail the survival requirement. These are unweighted counts over enumerated first actions, not policy-level episode frequencies. No rollouts are discarded, and the terminal flags permit verification that success equals target-first-wood acquisition and survival.

P MATCHED HISTORY INTERVENTIONS IN A TRAINED CALIBRATION

A second CPU calibration trains 10 deterministic GRU policies on one four-frame task with independent past bits x and k. A first action is rewarded for matching x. Only afterward a mode selects whether the final action should match past x, current k, or past k. We enumerate both fixed first actions and compare past-cue flips under identical current inputs. The current-x-absent stratum is prespecified. The payoff tests first or final correctness with equal probability.

Table 37: Means over all ten seeds. The ridge probe is fit on separate nuisance worlds and reads the final actor state. TV compares final action distributions after flipping past x. $D_{H,x}$ and $D_{H,k}$ square exact payoff differences after the respective past-cue flip, with uniform first-action weights. These history interventions differ from the current-cue D_w.

Final target	Probe x (%)	TV$_x$ (%)	$D_{H,x}$	$D_{H,k}$
Past x	100.00	98.115	0.241	2.08e-06
Current k	100.00	0.226	1.94e-06	0.027
Past k	100.00	0.275	3.01e-06	0.242

Table 37 separates retained information from the finite action response and its payoff effect. The prespecified practical checks and all individual-seed results are included with the calibration script and weights. Storage, local input sensitivity, finite action response, and payoff effects answer different questions; nonzero gradients are not classified as diagnostic errors. This deterministic small-task calibration does not test posterior coupling or establish diagnostic superiority in Dreamer.

On the same states, past-x-region recall shares are 8.89/5.09/6.92% and absolute pixel sensitivities are 0.000393/3.74e-05/0.000113 in mode order. Both gradient magnitudes and relative shares are reported; neither is a prediction of reward benefit. The two ignore-x modes have small, nonzero responses, satisfying prespecified tolerances (mean action TV $\leq 1\%$, payoff RMS difference ≤ 1 percentage point). The past-k mode supplies a relevant-history positive control. Past k is redundant in the current-k mode; conflicting past and current k can still change its final behavior, as the table reports. A positive first-action history-source gain remains in every mode, because the first reward still depends on x.

At the fixed update-300 secondary checkpoint, the past-k mode achieves 89.35% return and has 35.08% final-action TV under the past-x flip, with $D_{H,x} = 0.0828$. Its signed payoff contrast is zero to numerical precision: flipping the balanced irrelevant bit permutes the evaluated states, so positive and negative local effects cancel. This is endpoint sensitivity of an imperfect policy, not evidence that remembering x benefits the final-k decision.

Q DEVELOPMENT EVIDENCE AND NEGATIVE FINDINGS

The separate development supplement (`world-model-development-supplement.pdf`) retains the complete protocols, figures, tables, and selected examples for the earlier studies summarized here. Their exploratory status and all reported outcomes remain unchanged. These studies do not add independent training replicates to either principal matrix cohort.

Q.1 EXPLORATORY BASELINE PROVENANCE

The earlier roughly 500,000-step checkpoints predate explicit training-world seed recording; their training trajectories cannot be reconstructed. The observed 6.83-point difference between two Craftax configurations is not a robust configuration-superiority result. Saved counters range from 488,390 to 499,770. The label `size200m` does not itself establish parameter-count equivalence

with the published implementation. Full exploratory baseline values and provenance limitations remain in Supplement Section 1. The main manuscript separately reports the completed five-seed baseline evaluation.

Q.2 Immediate cues and one-transition development

The seed-0 one-transition pilot has 78 anchors. Recent history improves first turn by 6.57 points [3.37, 10.26], but its target-success contrast is only 1.76 points $[-0.16, 3.69]$. Recent and immediate cues differ in viewpoint as well as timing; their comparison does not isolate forgetting. The separate immediate-cue study finds a four-agent first-turn mean of 8.61 points, with a model-level t interval $[-2.42, 19.63]$. Continued visibility adds 1.62 target-success points over one true cue with the same first action. This is a supporting visibility comparison, not a direct comparison of the matrix's history-source gains. Supplement Sections 2–3 retain all condition-level outcomes and the original development/replication distinction.

Q.3 Common-state receiver differences

Across 384 donor anchors, receiver 4's recent-history first-turn gain differs from receivers 0–3 by -5.33 points $[-6.80, -3.99]$ under matched inputs. It retains immediate-cue and mismatch responses, while its ongoing-visible target success is 11.18 points lower $[-16.87, -5.55$ for the signed gap]. State selection alone is insufficient to explain the history-response gap, but visible competence remains a competing explanation. Supplement Sections 4–5 retain the complete 76,800-rollout panel and replay checks.

Q.4 Selected representation and intervention cases

On 45 held-out probe worlds, decision readouts for receivers 2 and 4 achieve 77.10% and 81.43%, while their history first-turn gains are 4.91 and -0.95 points. The readout difference includes zero; these results identify neither superior retention nor non-use. Target-directed first-actor edits produce a small target-success increase in receiver 2 and an unresolved increase in receiver 4. Neither native-return gain is resolved. Selected receivers, reused development worlds, task-labeled directions, and the absence of behavioral random-direction controls preclude autonomous memory-repair claims. Supplement Sections 6–9 preserve the readouts, perturbation screens, executed outcomes, and limitations.

Q.5 The source-action development null

The seed-0 development block retains all 77 anchors and has an unresolved free-minus-fixed interaction of -0.97 points $[-4.38, 2.44]$ at the 16-action endpoint. Its positive immediate first-turn cue effect does not yield a resolved final target benefit. The prior null remains separate from the unchanged four-agent replication in the main manuscript. Full results are in Supplement Section 10.

Q.6 Opposing individual interactions

Two selected development worlds have interactions of $+0.625$ and -0.625. In the latter, true and erased free first actions agree, yet the fixed action changes their subsequent outcome contrast. Interaction is therefore not equivalent to a cue-induced first-action change. Early acquisition advantages can also disappear as other conditions catch up. These selected examples, complete eight-draw counts, and the full-block event partition are retained in Supplement Section 12; they are not independent replications.

Q.7 Absolute conditions in the source-action replication

The six equal-agent condition means below retain the full absolute-outcome table from the earlier four-agent source-action replication. The fixed source has unedited visual history, so a larger free-minus-fixed cue contrast does not imply that free execution is a better controller.

Table 38: Absolute outcomes in the original source-action replication. Target penalty assigns every miss 17 actions; all rollouts in this block survive. The complete interpretation is preserved in Supplement Section 11.

Condition	Target (%)	Any wood (%)	Native return	Target penalty
Fixed, true	72.54	90.90	2.407	6.957
Fixed, erased	72.27	90.83	2.376	7.001
Fixed, mismatched	70.11	89.68	2.311	7.346
Free, true	71.60	89.53	2.380	7.312
Free, erased	66.42	88.07	2.326	8.206
Free, mismatched	52.44	83.86	2.238	10.437

Q.8 Prediction and edited-task diagnostics

Small edited-scene readouts have nonmonotone accuracy over delays and do not identify a forgetting curve. Seen history has higher native forecast MSE than erasure at every tested delay; continuous visibility also fails to improve these forecasts. Persistence beats the world-model forecast in every tested natural crop class, and swapped-content alignment is not positive in both natural stone and tree strata. In an edited two-action collection task, autonomous success is 10.55%, versus 98.44% when the correct first turn is supplied; autonomous success rises with a longer deadline. These early checks motivate the natural behavioral endpoint, without establishing a predictive-memory or general-control benefit. Supplement Section 13 retains all prediction scores and edited-task results.

Q.9 NOOP delay and unsuccessful restoration

Across the same 94 anchors, the seen-minus-erased target-success contrasts at delays 0, 2, 8, and 16 are 9.56, 8.56, 5.40, and 7.87 points. Visible success also falls from 81.28% to 43.59%, so this nonmonotone curve does not isolate retention. At delay 16, full-history restoration is worse than seen–masked input by 10.25 points [-16.57, -4.64 for the signed contrast]. Fresh true-versus-swapped cue effects remain unresolved at positive delays, as does the environment-reward return contrast at delay 16. Restoration is not a successful rescue, and these controls do not identify stored-but-unused information. Supplement Section 14 retains all delay conditions, pointwise development intervals, and behavioral partitions.

Q.10 Boundary-cue execution diagnostic

A separate 128-episode diagnostic asked whether a frozen actor that failed the delayed multi-command assay could use a newly exposed current-command panel at the first executable boundary. The actor and physical state were unchanged, the recurrent carry was retained, and the panel was shown for one observation only. This was a free, privileged, untrained intervention on reused development worlds; it is therefore a conditional execution diagnosis rather than a new history-source estimate or an independent model replicate.

At delay 8, first-command verification increased from 10/64 (15.6%) with no panel exposure to 19/64 (29.7%) after the one-frame pulse, a paired world bootstrap difference of +14.1 points [3.1, 25.0]. At delay 0 the corresponding difference was +4.7 points [-10.9, 20.3]. Persistent visibility produced 61/64 and 56/64 verified first commands at delays 8 and 0, respectively, whereas the pulse produced only 1/64 and 0/64 complete four-command episodes. Thus the actor can exploit a fresh cue in some delayed worlds but does not sustain the sequence after that cue disappears. The result does not identify forgetting, validate paid replay, estimate G, or establish a diagnostic decision benefit; full protocol and ledgers are in Supplement Section 15.

R Cohort-2 protected historical-source controls

This exploratory extension retains all 215 cohort-2 anchors (31/39/51/65/29 for seeds 5–9). Six sources replay the actual actions and rewards, observe the same true current image, and weight the original paired-draw Q_T and Q_E. No new trained models or continuation tails enter this extension.

Candidate black and candidate grass replace the two fixed candidate locations; near black and near grass use the other two axial neighbors of the decision anchor, fixed in world coordinates. All four protect player/entity tiles. Clean and candidate-swap sources complete the six conditions. Grass is rendered from a state copy with unchanged lighting and night-noise state; only selected tile pixels are copied. It is legal but potentially misleading terrain information.

Figure 8 shows the first anchor in the prespecified smoke panel (seed 5, reset-world index 5), without selecting on the source effects.

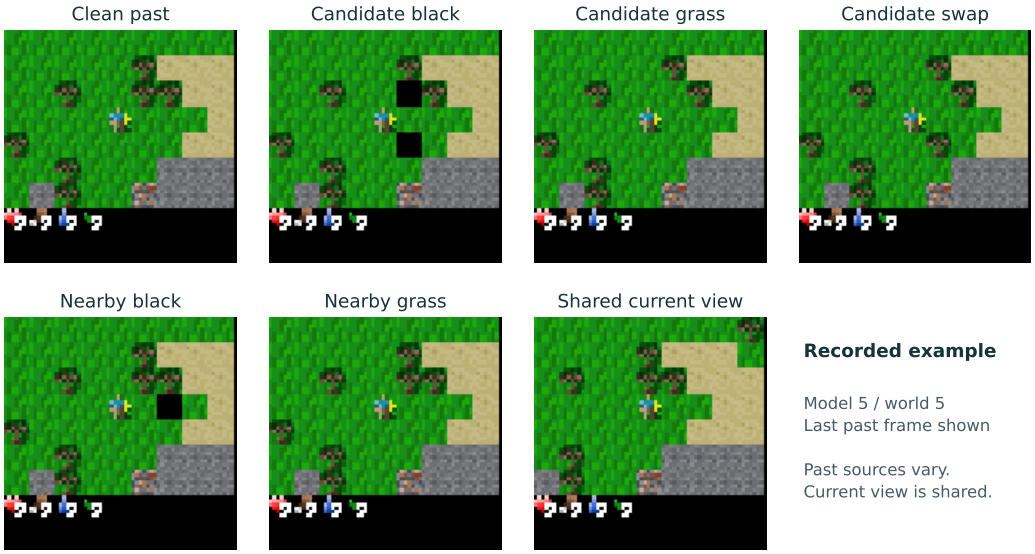

Figure 8: The final past frame from the fixed illustrative anchor under the six historical sources, followed by the shared true current image. Near grass equals clean in this example; the current image is unchanged across all sources. Full-prefix inference and all 215 anchors enter the reported estimates.

Table 39 reports the three contrasts specified before collection, all under Q_T. Models are weighted equally. The world intervals condition on the five checkpoints, while joint intervals additionally resample models. Both use 10,000 joint reset-world resamples; only five training units limit the approximation.

Table 39: Cohort-2 historical-source contrasts under Q_T, in target-success percentage points. The 98.33% intervals apply the three-contrast Bonferroni adjustment. All sources share a true current image; the saved continuations still mask candidates.

Contrast	Mean	World 95%	Joint 95%	World 98.33%	Joint 98.33%
Clean − candidate grass	+18.92	[14.59, 23.60]	[10.29, 29.43]	[13.67, 24.81]	[8.54, 32.25]
Near black − candidate black	-1.85	[-5.18, 1.34]	[-7.81, 3.87]	[-6.09, 2.15]	[-9.70, 5.35]
Near grass − candidate grass	+18.85	[14.52, 23.52]	[10.21, 29.41]	[13.63, 24.71]	[8.45, 32.00]

The corresponding joint adjusted intervals classify the three contrasts, in table order, as positive after the three-contrast adjustment; unresolved after the three-contrast adjustment; positive after the three-contrast adjustment. An unresolved contrast does not establish equivalence. These source controls do not by themselves isolate semantic memory or establish benefit under unmasked continuations.

Numerical bridge. A separate legacy-black inference preserves the original player-only protection rule. The maximum clean/legacy probability-component drift is 0; equal-model mean TV is 0 for clean and 0 for legacy black. Re-extracted minus archived history gain is 0.0000 [0.0000, 0.0000] pp under Q_T and 0.0000 [0.0000, 0.0000] pp under Q_E (joint 95% intervals). Protected

candidate-black minus legacy-black source value is +0.0152 [0.0000, 0.0862] / -0.0004 [-0.0020, 0.0000] pp under Q_T/Q_E; this changes the protection protocol and is not numerical drift.

Spatial limitations. Near grass changes only 33/1857 past frames and 6/215 anchor prefixes. It changes 1617/124999 selected pixels, versus 67267/159446 for candidate grass; clean-versus-near-grass action TV is 0.00069633. Its large contrast with candidate grass therefore largely repeats the clean contrast and does not establish matched-strength spatial specificity. Candidate and near support areas differ on 787/1857 past frames; their mean distance to the player differs on 512/1649 frames where both are editable. Total overwritten support is 159446 versus 124999 pixel locations over time. Entity protection changes the legacy-mask image on 4/1857 frames. These pooled geometry counts describe residual mismatch; near controls are neither perfectly spatially matched nor guaranteed irrelevant to the policy. Legal grass can still introduce false passability and tree-appearance cues.

S PAIRED CONTINUATION-VISIBILITY EXTENSION

This extension was specified before its collection, after the September 23 review. It selects the first 12 eligible world indices for each new-cohort model: 60 anchors over 33 distinct reset worlds. Each masked/unmasked pair starts from the same sampled recipient, simulator state, forced first action, posterior counter and tail random stream. Only observations after the forced action differ. Both visibility conditions are freshly executed in the same process; the selected archived masked values are a separate baseline.

Table 40: Equal-model means and joint model/world 95% bootstrap intervals. This selected panel does not estimate the full 215-anchor mean.

Continuation	G_T (pp)	G_E (pp)
Archived masked	+8.69 [+0.64, +17.35]	+7.96 [-1.28, +17.42]
Fresh masked	+8.71 [+0.67, +17.36]	+7.99 [-1.21, +17.46]
Fresh unmasked	+8.09 [-1.26, +17.05]	+8.01 [-1.53, +17.27]
Unmasked minus masked	-0.62 [-4.17, +2.08]	+0.02 [-3.21, +3.17]

The two extension-primary visibility changes have 97.5% Bonferroni intervals -0.62 [-4.95, +2.56] and +0.02 [-3.94, +3.71] pp (estimate followed by interval). The analysis resamples the union of selected reset-world IDs, preserving per-model selection masks, and resamples models for the joint intervals (10,000 replicates, seed 20260923). Conditional-world intervals and every model estimate are included in the reproduction archive. All 65,280 tail executions and 1,044,478 actual simulator transitions are retained, including 9 masked and 2 unmasked terminal tails. This extension is not an independent training replication; zero-spanning intervals do not establish equivalence.

Secondary outcomes and replay bridge. Environment-reward history-source gains under unmasked continuations are G_T: +0.0118 [-0.1050, +0.1249]; G_E: +0.0773 [-0.0697, +0.2458]. The fresh-minus-archived masked target gains are +0.02 [-0.00, +0.07] and +0.03 [-0.00, +0.11] pp. Mean source-probability TV drift (clean/T/E) is 0.0360/0.0337/0.0422%. These checks characterize the bridge to the archive; the within-process visibility comparison remains the extension target. All secondary intervals are exploratory and are not multiplicity controlled.

Interpretation. The selected-panel point estimates of history-source gain remain positive under both visibility conditions. However, both extension-primary adjusted intervals span zero, and both unmasked gain levels have zero-spanning joint 95% intervals. Thus the panel neither establishes equivalence across visibility conditions nor independently confirms unmasked history utility. It does not replace the original full-cohort masked-continuation findings.