WHEN THE DISPLAY IS WRONG: CONCEPT-DEPENDENT COMPENSATION IN WORLD-MODEL AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

A world-model agent may retain task information when its observations become misleading, but whether that information sustains control requires a causal test. In Craftax-Classic, we falsify inventory counts without changing actual resources. We compare four Δ-IRIS agents, whose policies receive frames, with ten DreamerV3 agents, whose policies receive recurrent and stochastic world-model states; all were trained with visible inventory. A false zero stone count reduces stone-related achievement rates from 98.9% to 3.3% in Δ-IRIS, but only from 97.9% to 93.8% in DreamerV3. This preserved performance motivates a within-agent test of recurrent support. Removing a readout-defined recurrent stone subspace has little effect with an accurate display but substantially increases impairment when the display is false. Replacing that component with one from a recorded donor state largely restores performance when the donor's resource status matches the agent's actual inventory; wrong-status donors do not. The correct-donor advantage grows from 2.6 percentage points with an accurate display to 26.3 with a false display, an increase of 23.8 points (Holm-corrected $p = 0.012$). Wood interventions are individually costly, and newly trained probes still decode resource availability after projection, limiting generalisation across resources and ruling out complete information erasure. However, correct- and wrong-status patches differ both in the magnitude of the recurrent-state change they induce and in their contextual similarity to the receiver, so restoration cannot be attributed solely to resource information. The results support partial recurrent compensation for misleading sensory input in the tested DreamerV3 agents, while the wood comparison shows that this capacity cannot be assumed for every resource represented by the same world model.

1 INTRODUCTION

An agent has just collected stone. Its inventory contains the resource, the simulator permits it to place stone, but the count on its display now says zero. This intervention creates a conflict between the world and a visible cue without changing what the agent can do. In Craftax-Classic, it nearly abolishes stone-related achievement in Δ-IRIS, while DreamerV3 remains largely successful. The contrast raises a mechanistic question: what supports action when the current observation reports the wrong resource status?

The policy interfaces make recurrent state a candidate: Δ-IRIS acts on observed or decoded frames, whereas DreamerV3 reads stochastic and recurrent world-model states (Micheli et al., 2024; Hafner et al., 2025). Architecture and training also differ, so we test recurrent support within DreamerV3. The pre-registered architecture experiment distinguished concept decodability, influence on next-code predictions and influence on imagined behaviour (App. A). It motivated a redundancy hypothesis: an alternative source may preserve behaviour after an internal intervention. We now test that possibility in the real environment.

Preserved performance alone cannot answer that question. A representation may be readable yet unused (Belinkov, 2022; Ravichander et al., 2021; Elazar et al., 2021). Conversely, a small lesion effect can conceal a contribution that another source makes unnecessary: self-repair and conditional

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.

co-ablation studies establish this conditional importance (McGrath et al., 2023; Rushing & Nanda, 2024; Gong et al., 2026). We therefore cross recurrent stone projection with display falsification. Removing the component becomes substantially more costly when the displayed count is false. This interaction identifies a conditional behavioural role, but increased sensitivity to generic disturbance remains an alternative explanation (Makelov et al., 2024; Li & Janson, 2024).

Donor restoration tests whether a recorded state with correct resource status helps more than one with wrong status. That advantage grows under the false display. Wood interventions, residual decodability and unequal donor patches then delimit the account: the results support partial recurrent compensation without isolating resource information as its cause. Dated protocols and the registry preserve the actual chronology, including exploratory follow-ups.

Our contributions connect three levels of evidence:

- **From display dependence to recurrent support.** The cross-agent contrast motivates a within-DreamerV3 test: recurrent stone projection reduces success by 0.006 with an accurate display and by 0.273 with a falsified one.
- **From joint impairment to donor-status sensitivity.** Correct-status patches outperform wrong-status patches by 0.026 with an accurate display and 0.263 with a falsified one. The resulting increase in the correct-donor advantage is similar across three donor-selection seeds ($+0.238$, $+0.248$ and $+0.245$). However, correct- and wrong-status patches differ in size and also alter non-target representations, so this advantage cannot be attributed solely to resource status.
- **From restoration to its mechanistic limits.** Probes trained on projected recurrent states without donor patches still decode resource availability (AUROC 0.978 for wood, 0.997 for stone), showing that projection does not erase all resource information. Routing the patched state separately to the actor or the transition model improves wood-tier success by 0.198 and 0.121, respectively, supporting contributions through both action selection and subsequent latent-state updates. However, the stochastic-repair control fails to establish effective repair, and policy retraining remains inconclusive.

2 Study design and causal questions

2.1 Agents, observations and outcomes

We study Craftax-Classic, which provides pixels, inventory variables and simulator-scored achievements (Matthews et al., 2024). We distinguish *HUD visibility during training* from *displayed-count accuracy during evaluation*. Agents in the *HUD-visible* (rendered) training condition see the inventory heads-up display; those in the *HUD-hidden* condition have the entire HUD, including vitals, masked while the world view is unchanged (Tao et al., 2025). These observation conditions are retained during evaluation. Our main test-time intervention applies to HUD-visible agents: the HUD remains on screen, but the target resource count is either *accurate* (the true count) or *falsified* (forced to zero). Falsification changes one displayed count, leaving the simulator's inventory intact; it does not hide the HUD. It is a semantically targeted observation perturbation, related to the state-adversarial formulation of Zhang et al. (2020), rather than a change to the environment's resource state. A sham edit redraws the correct count using the same pixel-replacement procedure, controlling for visual artifacts. As a specificity control, we falsify the other resource's cell while leaving the target resource's display accurate.

DreamerV3's recurrent state-space model (RSSM) has recurrent state $h_t \in \mathbb{R}^{1024}$ and stochastic state z_t (32 one-hot categoricals of 32 classes); its actor reads $[z_t, h_t]$. Δ-IRIS predicts discrete transition codes and decodes frames for imagined control, but acts directly on observed frames in the real environment. World-model lesions therefore do not directly change its real policy input; our recurrent interventions apply only to DreamerV3 (Fig. 1).

The architecture experiment trained 37 agents for approximately 10^7 environment steps each, plus 20 matched-budget controls to test unequal training exposure (App. A). Before intervention evaluation, the imagined-competence screen selected 17 DreamerV3 agents (10 HUD-visible, 7 HUD-hidden) and four HUD-visible Δ-IRIS agents. Results concern this selected cohort. The ten HUD-visible DreamerV3 agents support the main compensation test because both their familiar inventory cue and

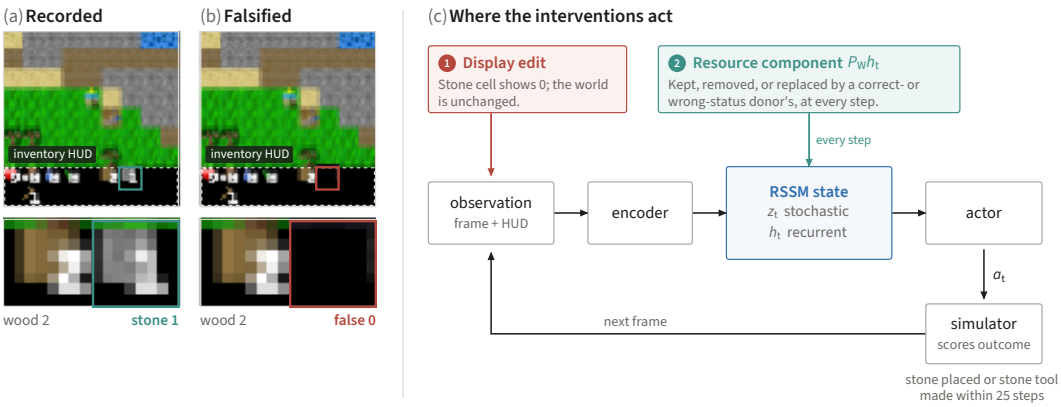


Figure 1: **Falsified display and intervention sites. (a)** Recorded frame just after a DreamerV3 agent trained with the HUD visible collects stone; the boxed HUD cell shows one stone (enlarged below). **(b)** The same frame with that cell replaced by a recorded cell showing no stone; the simulator's inventory is unchanged. **(c)** The display edit (1) acts on the observation; the recurrent interventions (2) act on the resource component $P_W h_t$ at every step. The simulator scores stone placement or a stone tool within 25 steps.

recurrent policy input can be manipulated. The four Δ-IRIS agents provide the display-falsification comparison. App. M links experiments F1–F7 to protocols and results.

To test resource use after acquisition, rollouts start from recorded first-collection states with the corresponding resource-tier achievement still pending. Success is recorded by the simulator, avoiding reliance on a learned achievement detector. Wood-tier success includes placing a table or making a wooden tool; stone-tier success includes placing stone or making a stone tool. Most successes are table or stone placement, respectively, so these endpoints measure downstream resource use rather than tool-making alone.

The evaluation window balances time to succeed against sensitivity to delays. A pilot selected 25 steps for HUD experiments as the shortest tested window with at least 80% intact success; longer windows approached ceiling. Latent repair retains the pre-specified 13-step F3 window, with room for both impairment and restoration in wood. Stone uses the same window, but its near-ceiling baseline limits the restoration control (App. I).

Conditions share a hash-verified simulator starting state and matched rollout randomness, reducing variation from initial conditions and stochastic events. We average success within each agent, then across agents, giving trained agents equal weight and treating them as the independent units. Effects are absolute probability differences: 0.01 is one percentage point.

2.2 State interventions and content controls

For each agent and resource, logistic readouts are fitted on design episodes to predict resource availability and whether specified items can be crafted. Their coefficient directions in raw-state coordinates span an orthonormal basis W: rank 4 for wood and rank 3 for stone after duplicate labels are removed. Let $P_W = WW^\top$; we call $P_W h$ the resource component of h. We either remove it or replace it with the corresponding component of a recorded donor state:

$$h^{\text{drop}} = (I - P_W)h, \qquad h^{\text{donor}} = (I - P_W)h + P_W h_{\text{donor}}. \tag{1}$$

These operations remove or replace directions; they do not by construction erase every encoding of the resource. Following the distinction between decodability and behavioural use in amnesic probing (Elazar et al., 2021), independent readouts fitted on disjoint evaluation episodes test what remains. The validation loads the exact deployed basis and stochastic slots, rather than fitting a similar intervention again.

The *receiver* is the state being patched. A *correct-status* donor matches a fixed key of availability, craftability, placeability, table-proximity and selected inventory flags; App. I lists both resource-

specific keys. A *wrong-status* donor has opposite resource availability, with dependent flags recomputed and the remaining key flags matched. If no exact match exists, criteria are relaxed in a fixed, logged order. All conditions use the same rule and random seeds, but their donor identities can differ once trajectories diverge. In the stone restoration experiment, exact key matches occur in approximately 99.98% of correct-status and 99.99% of wrong-status patches (means over agents).

The donor key does not explicitly match whether the receiver faces a wall. Exact correct-status matching fixes this flag only with at least one stone or two wood; wrong-status matching gives no such guarantee. Offline stone wall-status agreement is 86% for correct and 73% for wrong donors (medians over ten F4R agents; descriptive rates, not a chance test). Matching also does not ensure preservation of non-target representations, the concern evaluated by RAVEL (Huang et al., 2024); we test this in §4.1 and App. K.

The latent repair experiment also replaces eight selected stochastic slots with donor codes. After damaging both h and z, we repair h, z, or both. Positive controls test whether each repair reverses a lesion at that site alone. Recurrent repair is also routed separately to the actor or transition model. Lesions continue on subsequent steps; repair at the first step is compared with repair at every step. The experimenter supplies donor content using simulator labels, so restoration does not demonstrate autonomous reconstruction.

2.3 What counts as compensation?

Let $Y(f, r)$ be the mean simulator success when the interventions are applied at every rollout step: the display is accurate ($f = 0$) or shows a false zero ($f = 1$), and the recurrent component is kept, removed, or replaced by a correct- or wrong-status donor's (r). All conditions share their starting states and the specified rollout and donor randomness. Write $\Delta_H(f) = Y(f, \text{drop}) - Y(f, \text{keep})$ for the change in success when the component is removed; it is negative when removal hurts. The quantity we test is whether removal hurts more under the false display,

$$I = \Delta_H(1) - \Delta_H(0) = \big[Y(1, \text{drop}) - Y(1, \text{keep})\big] - \big[Y(0, \text{drop}) - Y(0, \text{keep})\big], \tag{2}$$

which is negative in that case. A negative I is compatible with compensation, but also with the component simply becoming more fragile once the display is wrong. Restoration therefore adds a content contrast:

$$\begin{aligned} S(f) &= Y(f, \text{correct}) - Y(f, \text{wrong}), \\ \Psi &= S(1) - S(0). \end{aligned} \tag{3}$$

A positive Ψ means that the advantage of correct over wrong content is larger under the false display. Attributing that advantage to resource content alone would further require patches of equal size and no change to other represented properties. Because each rollout's state keeps evolving after every patch, these conditions do not set two fixed variables independently. We use *partial compensation* for the pattern of joint impairment plus content-sensitive restoration, without claiming complete substitution or erasure of information. Self-repair and conditional co-ablation studies already show why interventions made with every other source intact can miss effects that appear only when another source is changed (McGrath et al., 2023; Gong et al., 2026; Chen et al., 2025). Our crossed design measures this dependence in an acting agent; its interpretation remains specific to the chosen outcome and interventions (Zhang & Nanda, 2024; Heimersheim & Nanda, 2024).

The architecture experiment was pre-registered at five seeds per architecture–HUD condition. Extensions to 7–12 seeds followed the initial tests and are reported as sequential. Follow-up and repair experiments had separate dated protocols committed before their runs, with questions motivated by the preceding results. Paired effects are averaged within each trained agent before two-sided Wilcoxon tests and Holm correction within specified contrast families. Equivalence is assessed using two one-sided tests (TOST) with a predeclared ± 0.05 success-rate margin; a nonsignificant difference alone is insufficient. Protocols, deviations and superseded results remain in the appendix.

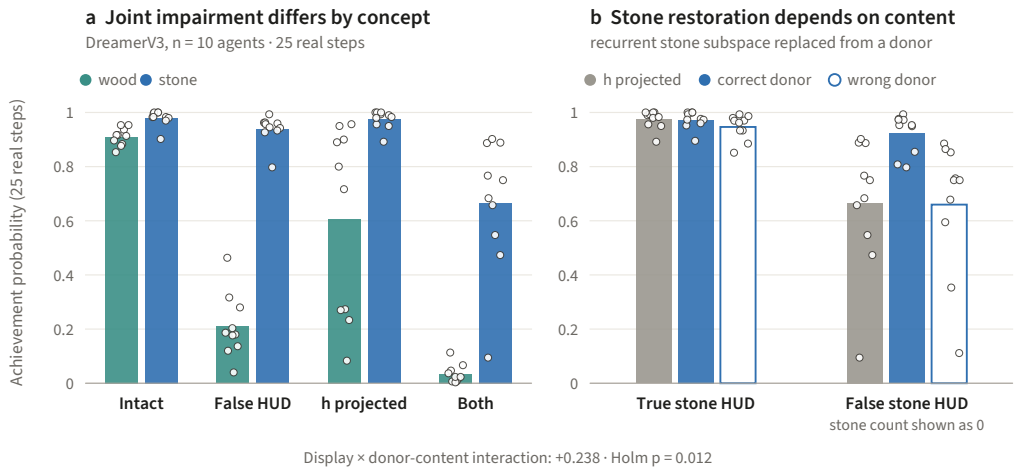


Figure 2: **Recurrent effects depend on displayed-count accuracy and resource identity.** Resource-tier success within 25 steps for ten DreamerV3 agents trained with the HUD visible. "True stone HUD" shows the actual stone count; "False HUD" ("False stone HUD" in b) forces only the target resource count to zero, leaving the simulator's inventory unchanged. **(a)** Count falsification and recurrent projection have larger individual effects for wood than for stone; for stone, the combined loss exceeds the sum of the individual losses. Achievement criteria and baselines differ across resources. **(b)** Under the falsified stone count, correct-status donor patches restore success whereas wrong-status patches do not (F4R). Bars show agent means; dots show individual agents. The correct-donor advantage is 0.026 with an accurate count and 0.263 with a falsified count; $\Psi = +0.238$ (Holm $p = 0.012$).

3 From display dependence to recurrent compensation

3.1 Preserved performance motivates a recurrent-state test

The display perturbation first establishes the behavioural contrast. With visible-inventory training, showing a false zero stone count lowers success from 0.989 to 0.033 in Δ-IRIS, but from 0.979 under sham redraw to 0.938 in DreamerV3. These are responses to a changed observation in the real environment. They motivate testing recurrent support within DreamerV3; they do not establish that the difference is caused by the policy interface.

If recurrent state supports the preserved behaviour, its contribution should increase when the display is false. With an unedited display, recurrent stone projection has little effect ($0.980 \rightarrow 0.974$). Combining projection with display falsification lowers success to 0.665 (Fig. 2a). The interaction in Eq. 2 is -0.268 and is negative in all ten agents (paired Wilcoxon $p = 0.002$ in the crossed experiment, F4; Holm $p = 0.008$ in its restoration extension, F4R). The small recurrent-only effect therefore conceals a much larger contribution under misleading input.

The false display alone still causes a small, statistically detectable loss. The joint loss could also reflect increased sensitivity to disturbance. Donor restoration tests whether performance depends on the resource status of the supplied state.

3.2 Correct-status donor patches restore performance under a falsified display

F4R (App. M, row F4R) repeats the display-by-recurrent design and adds correct- and wrong-status donor patches. The five conditions shared with F4 reproduce its recorded outcomes exactly. With the count still falsified, correct-status restoration raises success from 0.665 to 0.923 ($+0.258$; improvement in 9 of 10 agents; Holm $p = 0.012$). Wrong-status patches give 0.660; the correct-donor advantage is 0.263 (positive in 9 of 10 agents; Holm $p = 0.012$). This difference supports a role for supplied content, subject to the intervention limits below.

Table 1: **Display × donor-content comparison for stone.** Mean stone-tier success probability within 25 steps over ten agents (F4R, donor seed 0). Donor status is defined relative to the agent's actual inventory at each step. The correct-donor advantage increases under falsification: $\Psi = +0.238$ (Holm $p = 0.012$). The falsified-display, wrong-donor condition is lowest in 9 of 10 agents; the pattern recurs with two further donor seeds (App. J).

Displayed stone count	Correct-status donor	Wrong-status donor
Accurate	0.972	0.946
Falsified	0.923	0.660

These outcomes support partial compensation, but not a Boolean OR/AND computation: patches do not set all recurrent resource information to a binary value. Two additional donor seeds give restoration effects of $+0.260$ and $+0.257$ and correct-donor advantages of 0.272 and 0.270. All remain significant after correction, with per-agent rank correlations of at least 0.94 across seeds (App. J).

With an accurate display, correct and wrong patches yield 0.972 and 0.946 success; the 0.026 advantage is not detectable after correction (Table 1). Under falsification it grows to 0.263, giving $\Psi = +0.238$ from unrounded estimates (positive in 9 of 10 agents; Holm $p = 0.012$). This directly tests the change in donor effect. The interaction is positive and significant at all three donor seeds ($+0.238$, $+0.248$, $+0.245$), and primary contrasts also hold on the log-odds scale (App. J).

Patch size remains a confound. A pre-registered norm-matching attempt failed its 10% criterion (median errors 18.5% and 10.7%). Matching candidate patches at one state did not equalise the applied patches across diverging rollouts. Both arrangements retained positive content effects, but a post-hoc comparison of norm-selected conditions gave $+0.092$ ($p = 0.084$, uncorrected); full norms and contrasts are in App. J. These controls cannot separate resource content from perturbation size or other represented properties. Wall status is a concrete concern: donor agreement differs by status (§2), and its readout shows the largest non-target changes (§4.1).

3.3 Wood tests the scope of the compensation account

A second resource tests whether the stone pattern generalises within the same agents. For wood, each intervention reduces DreamerV3's success: the HUD edit leaves 0.210, recurrent projection 0.607, and both 0.034, against 0.910 intact. Success under the joint intervention is already near zero, leaving little room for further loss. The paired stone–wood difference in interactions is -0.394 ($p = 0.010$), but different achievement criteria, baselines and floor effects limit its interpretation. Wood interventions are individually disruptive, while the tested stone component has a much larger effect specifically under HUD falsification.

Wood HUD falsification impairs all 14 agents trained with the HUD visible ($0.928 \rightarrow 0.161$ pooled), although several retain substantial success. Evidence for recurrent compensation therefore concerns stone in the tested DreamerV3 agents; it is not a uniform response to inventory falsification.

4 Resource information remains decodable despite behavioural impairment

4.1 Projection does not establish erasure

The behavioural effects leave a further question: did projection erase resource information, or disrupt directions used by the existing policy? A fixed readout and a newly trained readout address different parts of that question. A wood-availability readout trained on intact states falls from AUROC 0.999 to 0.783 after projection. Newly trained readouts on projected states reach 0.978 (both logistic and MLP, median over 17 agents). For stone, the original readout retains 0.948 and a newly trained logistic readout reaches 0.997. Random projection of the same rank changes these readouts little. Resource information therefore remains accessible to a new decoder even where the existing policy's behaviour changes.

Table 2: **What the wood repair establishes.** Effects on simulator-scored wood-tier achievement within 13 steps, after both latent interventions (F3R2; 17 agents). Holm correction is within each predeclared family. A small z-repair effect cannot identify z's role because its positive control fails.

Repair condition	Change in success	Holm p
Recurrent component restored	$+0.314$	0.000061
Stochastic slots restored	$+0.016$	0.043
Both restored	$+0.317$	0.000061
Recurrent repair read by actor only	$+0.198$	0.000092
Recurrent repair passed to transition only	$+0.121$	0.0014
Recurrent repair at first step only	$+0.002$	0.286

A wrong-status recurrent patch moves an independent readout trained on intact states toward the donor's resource status (AUROC 0.85 for wood, 0.79 for stone); correct-status patches rarely flip its decisions. Wrong-status patches flip 77% of wood-availability decisions and 57% of stone-availability decisions, against mean non-target flip rates of 5% and 8%. The non-target readout that moves most is for the cell in front of the agent (placeability, one of the labels defining the basis, depends on it); it changes more under wrong- than correct-status patches: 18.7% versus 14.9% for wood and 20.2% versus 15.6% for stone (medians over 17 agents). For the ten agents trained with the HUD visible in F4R, the stone rates are 24.6% versus 16.2%. These checks show sensitivity to target status but do not exclude behavioural effects from changes to other represented properties. They use random donors on held-out states and do not validate the norm-selected donors (App. K).

4.2 Supplied wood content helps through more than one route

The wood experiments also test how supplied recurrent content reaches behaviour. In the 13-step repair experiment, damaging both the selected stochastic slots and recurrent subspace lowers success from 0.565 to 0.208. Restoring the recurrent component raises it to 0.521 ($+0.314$; improvement in all 17 agents; Holm $p = 6 \times 10^{-5}$). Donor-seed replications give similar effects ($+0.311$ and $+0.316$). In the positive controls, recurrent repair reverses a recurrent-only lesion, but stochastic repair does not reliably reverse a stochastic-only lesion.

Relative to damaging both sites without repair, routing the repaired recurrent state only to the actor improves success by 0.198; routing it only to the transition model improves success by 0.121 (Table 2). Transition repair can influence future stochastic and recurrent states. Repair only at the first step is insufficient while lesions continue, whereas repair at every step helps; intermediate schedules were not tested. Supplied donor content can therefore help through both routes, without establishing autonomous recovery or an exclusively direct h-to-action mechanism.

Stochastic repair improves success by 0.016 after both lesions, but by only 0.011 in its positive control with h intact ($p = 0.10$). Donor-code replacement changes an average of 7.1 of eight selected slots while changing independent availability readouts little. The weak repair effect therefore leaves z's role unresolved. The wood $z \times h$ lesion interaction is equivalent to zero within the predeclared margin (-0.006, 90% CI $[-0.048, +0.036]$, TOST $p = 0.042$), consistent with additive effects of these operations on this endpoint. It does not exclude other redundant encodings.

4.3 Relearning with restricted access remains inconclusive

Decodability after projection leaves open whether a policy can learn to use the remaining directions. To test adaptation, a fresh actor–critic is trained in each frozen world model under three input conditions: the recurrent wood subspace removed, a random subspace of equal rank removed, or no removal (F7). The world model retains its full state; training begins from recorded states of the existing agent. Initial actor–critic weights are paired across conditions, with three learner seeds. The no-removal learner passes the preset competence threshold in eight of ten HUD-visible agents.

Compared with no removal, wood masking changes success by -0.049 (95% CI $[-0.103, +0.005]$; Holm $p = 0.22$); equivalence within ± 0.05 is not established. Relative to random masking, success is lower by 0.050 ($p = 0.047$); excluding the agent used to calibrate the budget leaves a similar mean but $p = 0.094$. None of the seven HUD-hidden agents' no-removal learners passes the competence

threshold at the original budget, and neither of two extended-budget pilots passes it. Adaptation to restricted policy inputs therefore remains unresolved. This experiment does not remove the underlying information from the world model.

5 WHY THE BEHAVIOURAL ENDPOINT AND INTERVENTION CONTROLS MATTER

The preceding tests rely on simulator outcomes and validated intervention objects. The preregistered architecture experiment illustrates why those choices matter: it compared architectures trained with the HUD visible or hidden. Hiding it changes Δ-IRIS's detector-estimated wooden-tool-tier rate in imagined rollouts by -0.314 and DreamerV3's by $+0.004$; the interaction is $+0.318$ (permutation $p = 0.004$, 37 agents). Frame-only readouts lose about 0.111 AUROC in both architectures (difference 0.000, 95% CI $[-0.032, +0.032]$), while DreamerV3's recurrent state retains more readable inventory information. These results are consistent with different access to task information, but architecture, objectives, detector families and removal of visible vitals prevent attribution solely to the control interface (App. A).

The direction selected in that experiment (the *legacy gate*) produces a larger imagined cascade—a drop in detector-estimated achievement during world-model rollouts—when its lesion reaches h than when it is restricted to z. Routing the same lesion only to the actor lowers simulator-scored wood-tier success by 0.087; routing it only to the transition model lowers it by 0.100 (F1). So the intervention affects behaviour, but readouts show that the direction carries the wood-collection event, or little wood information at all, rather than wood availability. Its behavioural effect does not establish removal of wood-availability information.

Measurement adds another ambiguity. A lesion applied only to the detector changes the imagined achievement estimate. On recorded simulator trajectories, the detector's median positive rate over 17 agents is 0.566 on 12-frame windows without a tier event, versus 0.966 with an event. This diagnostic does not quantify detector error on imagined trajectories. The main compensation and restoration results instead use simulator-recorded achievements. Also, 79% of Δ-IRIS single-step (concept, seed) contrasts fail to exceed the variance-matched control threshold; that test concerns next-code prediction, not simulator behaviour (App. D).

6 RELATED WORK

Compensation and conditional importance. The Hydra effect and subsequent work establish partial, heterogeneous compensation after ablation (McGrath et al., 2023; Rushing & Nanda, 2024). Circuit-completeness analyses identify interactions missed by isolated lesions (Chen et al., 2025); Conditional Co-Ablation reveals dormant backups and validates them through freezing, completion and knockout tests (Gong et al., 2026). Conditional importance and causal validation thus precede our study. We test compensation between misleading observations and recurrent state during closed-loop action, with outcomes scored by the simulator. Donor restoration is supplied externally; it establishes neither autonomous reconstruction nor Boolean circuit types or an information-theoretic redundancy decomposition (Liang et al., 2023).

Decodability, erasure and intervention validity. Probing reveals accessible information, not necessarily its use (Hewitt & Liang, 2019; Ravichander et al., 2021; Belinkov, 2022); amnesic probing links removal to behaviour (Elazar et al., 2021). INLP, adversarial linear erasure and LEACE target specified decoder classes (Ravfogel et al., 2020; 2022; Belrose et al., 2023). Our projection has no erasure guarantee, and retrained probes recover availability. Causal abstraction tests internal variables (Geiger et al., 2021; 2025); DAS learns alignments by interchange (Geiger et al., 2024), whereas we select directions by supervised readouts. RAVEL tests target changes and preservation of other attributes (Huang et al., 2024); our matched donor flags offer narrower control. Patching can induce misleading representations, with conclusions depending on metrics, corruption and ablation baselines (Makelov et al., 2024; Grant et al., 2026; Zhang & Nanda, 2024; Heimersheim & Nanda, 2024; Li & Janson, 2024). Successful rescue does not remove these concerns.

Causal studies of agents and world models. World models support control through predicted observations or latent dynamics (Ha & Schmidhuber, 2018; Hafner et al., 2019; Micheli et al., 2024; Hafner et al., 2025); architecture and temporal context affect performance (Deng et al., 2023; Samsami et al., 2024). Probes and interventions reveal Othello board representations (Li et al., 2023; Nanda et al., 2023), chess look-ahead (Jenner et al., 2024) and recurrent Sokoban planning (Bush et al., 2025; Taufeeque et al., 2025). Closest to our setting, Zhang (2026) measures prediction changes after IRIS and DIAMOND latent interventions. We score environment achievements and cross recurrent interventions with display accuracy. General-agent theory (Richens et al., 2025) and tests of predictive-model coherence (Vafa et al., 2024) address world-model competence without identifying the resource directions used by a policy.

Misleading observations and concept interfaces. Our false inventory count is a targeted observation attack that preserves simulator state (Zhang et al., 2020). Cue dependence relates to shortcut learning (Geirhos et al., 2020), but failure under this edit does not establish pursuit of a wrong objective (Langosco et al., 2022). Concept bottlenecks make concepts explicit control variables (Koh et al., 2020; Ye et al., 2025; Kulkarni et al., 2025); our agents lack such bottlenecks, so semantic selectivity requires post-training validation.

Dictionary features and the scope of the claim. Sparse dictionaries expose candidate directions (Bricken et al., 2023; Cunningham et al., 2024; Gao et al., 2025), but reconstruction quality need not predict probing or benchmark utility (Kantamneni et al., 2025; Karvonen et al., 2025). Following causal claim discipline (Joshi et al., 2026), we distinguish decodable information, behaviourally consequential directions and donor content that helps under a misleading display.

7 Discussion

A false stone count nearly abolishes achievement in Δ-IRIS while leaving DreamerV3 largely successful. Within DreamerV3, a recurrent projection with little effect under an accurate display becomes costly under falsification, and correct-status donors largely reverse the joint impairment. Wood and readout tests delimit the explanation: dependence varies by resource, and behavioural impairment does not imply information erasure. The evidence supports partial recurrent compensation, while unequal patch sizes and context leave its content-specific mechanism unresolved.

Crossed interventions reveal conditional importance (McGrath et al., 2023; Gong et al., 2026), not necessity or interchangeability. Donor restoration does not demonstrate autonomous reconstruction, and a binary display edit establishes no reliability–reliance law. The stochastic-state mechanism and policy adaptation remain unresolved; architecture comparisons do not isolate the interface. Compensation must therefore be assessed per resource, policy and outcome.

AI use statement

The experiment code, cluster execution, analyses, and figures were developed with an AI coding agent (Claude). Codex assisted with independent audits, literature verification and restructuring the manuscript. The authors checked the experiment and analysis code and are responsible for the research questions, protocol decisions, interpretation, and verification of the reported results. Dated protocols and the per-run provenance index document the available records and their limitations (App. M).

Reproducibility statement

Initial predictions are recorded at tag `prereg-v1`; dated protocols govern the follow-ups. Code, configurations, aggregate results and figure sources accompany the paper. App. K records corrections and protocol deviations; App. N documents source-tracking gaps, including missing generating commits and unrecorded working-tree changes. Instrumented reruns reproduce the outcomes they re-execute, but do not reconstruct all historical source states.

REFERENCES

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 04 2022. ISSN 0891-2017. doi: 10.1162/coli_a_00422. URL https://doi.org/10.1162/coli_a_00422.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/d066d21c619d0a78c5b557fa3291a8f4-Paper-Conference.pdf.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

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

Hang Chen, Jiaying Zhu, Xinyu Yang, and Wenya Wang. Rethinking circuit completeness in language models: And, or, and adder gates. In *Advances in Neural Information Processing Systems*, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/dd37fdb24a4e1cfa3ed5c247217a7394-Paper-Conference.pdf.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations (ICLR)*, 2024.

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://papers.neurips.cc/paper_files/paper/2023/hash/e6c65eb9b56719c1aa45ff73874de317-Abstract-Conference.html.

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

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *International Conference on Learning Representations (ICLR)*, 2025.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. 2021. URL https://arxiv.org/abs/2106.02997.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, pp. 160–187, 2024. URL https://proceedings.mlr.press/v236/geiger24a.html.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 2025. URL https://www.jmlr.org/papers/v26/23-0058.html.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2020. URL https://doi.org/10.1038/s42256-020-00257-z.

Zhiren Gong, He Lu, Tiantong Wang, Yichi Zhang, Yixin Wang, Zihao Zeng, Ming Xiao, Chau Yuen, and Wei Yang Bryan Lim. Conditional co-ablation: Recovering self-repair backups in transformer circuits. *arXiv preprint arXiv:2607.01940*, 2026. URL https://arxiv.org/abs/2607.01940v2.

Satchel Grant, Simon Jerome Han, Alexa Tartaglini, and Christopher Potts. Addressing divergent representations from causal interventions on neural networks. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/133e588e1429f9f1e25b215da145580e-Abstract-Conference.html.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. 2019. URL https://arxiv.org/abs/1811.04551.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, 2025.

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*, 2024. URL https://arxiv.org/abs/2404.15255.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. 2019. URL https://arxiv.org/abs/1909.03368.

Jing Huang, Zhengxuan Wu, Christopher Potts, Mor Geva, and Atticus Geiger. RAVEL: Evaluating interpretability methods on disentangling language model representations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8669–8687, 2024. URL https://aclanthology.org/2024.acl-long.470/.

Erik Jenner, Shreyas Kapur, Vasil Georgiev, Cameron Allen, Scott Emmons, and Stuart Russell. Evidence of learned look-ahead in a chess-playing neural network. In *Advances in Neural Information Processing Systems*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/37d9f19150fce07bced2a81fc87d47a6-Abstract-Conference.html.

Shruti Joshi, Aaron Mueller, David Klindt, Wieland Brendel, Patrik Reizinger, and Dhanya Sridhar. Position: Causality is key for interpretability claims to generalise. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://openreview.net/pdf/148d0ec00d1f63f98fa3f95f680199d47b61753e.pdf. Position paper.

Subhash Kantamneni, Joshua Engels, Senthooran Rajamanoharan, Max Tegmark, and Neel Nanda. Are sparse autoencoders useful? A case study in sparse probing. In *International Conference on Machine Learning (ICML)*, 2025.

Adam Karvonen, Can Rager, Johnny Lin, Curt Tigges, Joseph Isaac Bloom, David Chanin, Yeu-Tong Lau, Eoin Farrell, Callum Stuart Mcdougall, Kola Ayonrinde, Demian Till, Matthew Wearden, Arthur Conmy, Samuel Marks, and Neel Nanda. SAEBench: A comprehensive benchmark for sparse autoencoders in language model interpretability. In *International Conference on Machine Learning*, 2025. URL https://proceedings.mlr.press/v267/karvonen25a.html.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning*, 2020. URL https://proceedings.mlr.press/v119/koh20a.html.

Akshay Kulkarni, Ge Yan, Chung-En Sun, Tuomas Oikarinen, and Tsui-Wei Weng. Interpretable generative models through post-hoc concept bottlenecks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Kulkarni_Interpretable_Generative_Models_through_Post-hoc_Concept_Bottlenecks_CVPR_2025_paper.html.

Lauro Langosco Di Langosco, Jack Koch, Lee D Sharkey, Jacob Pfau, and David Krueger. Goal misgeneralization in deep reinforcement learning. In *International Conference on Machine Learning*, 2022. URL https://proceedings.mlr.press/v162/langosco22a.html.

Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *International Conference on Learning Representations (ICLR)*, 2023.

Maximilian Li and Lucas Janson. Optimal ablation for interpretability. In *Advances in Neural Information Processing Systems*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/c55e6792923cc16fd6ed5c3f672420a5-Abstract-Conference.html.

Paul Pu Liang, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, Richard Chen, Zihao Deng, Nicholas Allen, Randy Auerbach, Faisal Mahmood, Ruslan Salakhutdinov, and Louis-Philippe Morency. Quantifying & modeling multimodal interactions: An information decomposition framework. In *Advances in Neural Information Processing Systems*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/575286a73f238b6516ce0467d67eadb2-Abstract-Conference.html.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. Is this the subspace you are looking for? An interpretability illusion for subspace activation patching. In *International Conference on Learning Representations (ICLR)*, 2024.

Michael Matthews, Michael Beukman, Benjamin Ellis, Mikayel Samvelyan, Matthew Jackson, Samuel Coward, and Jakob Foerster. Craftax: A lightning-fast benchmark for open-ended reinforcement learning. 2024. URL https://arxiv.org/abs/2402.16801.

Thomas McGrath, Matthew Rahtz, Janos Kramar, Vladimir Mikulik, and Shane Legg. The Hydra effect: Emergent self-repair in language model computations. *arXiv preprint arXiv:2307.15771*, 2023. URL https://arxiv.org/abs/2307.15771.

Vincent Micheli, Eloi Alonso, and François Fleuret. Efficient world models with context-aware tokenization. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Neel Nanda, Andrew Lee, and Martin Wattenberg. Emergent linear representations in world models of self-supervised sequence models. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 16–30, 2023.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Annual Meeting of the Association for Computational Linguistics*, 2020. URL https://aclanthology.org/2020.acl-main.647/.

Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan D Cotterell. Linear adversarial concept erasure. In *International Conference on Machine Learning*, 2022. URL https://proceedings.mlr.press/v162/ravfogel22a.html.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. Probing the probing paradigm: Does probing accuracy entail task relevance? In *European Chapter of the Association for Computational Linguistics (EACL)*, 2021.

Jonathan Richens, David Abel, Alexis Bellot, and Tom Everitt. General agents contain world models. In *International Conference on Machine Learning*, 2025. URL https://arxiv.org/abs/2506.01622.

Cody Rushing and Neel Nanda. Explorations of self-repair in language models. In *International Conference on Machine Learning*, 2024. URL https://proceedings.mlr.press/v235/rushing24a.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://proceedings.iclr.cc/paper_files/paper/2024/hash/0a0e436e8830f746300f592377730fca-Abstract-Conference.html.

Ruo Yu Tao, Kaicheng Guo, Cameron Allen, and George Konidaris. Benchmarking partial observability in reinforcement learning with a suite of memory-improvable domains. *Reinforcement Learning Journal*, 6:1412–1439, 2025.

Mohammad Taufeeque, Philip Quirke, Maximilian Li, Chris Cundy, Aaron David Tucker, Adam Gleave, and Adrià Garriga-Alonso. Planning in a recurrent neural network that plays Sokoban. *arXiv preprint arXiv:2407.15421v3*, 2025. URL https://arxiv.org/abs/2407.15421v3. Expanded version of work presented at the ICML 2024 Mechanistic Interpretability Workshop.

Keyon Vafa, Justin Y. Chen, Ashesh Rambachan, Jon Kleinberg, and Sendhil Mullainathan. Evaluating the world model implicit in a generative model. In *Advances in Neural Information Processing Systems*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/2f6a6317bada76b26a4f61bb70a7db59-Abstract-Conference.html.

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

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/06a52a54c8ee03cd86771136bc91eb1f-Abstract-Conference.html.

Huan Zhang, Hongge Chen, Chaowei Xiao, Bo Li, Mingyan Liu, Duane Boning, and Cho-Jui Hsieh. Robust deep reinforcement learning against adversarial perturbations on state observations. In *Advances in Neural Information Processing Systems*, 2020. URL https://papers.nips.cc/paper_files/paper/2020/hash/f0eb6568ea114ba6e293f903c34d7488-Abstract.html.

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

A THE PRE-REGISTERED ARCHITECTURE EXPERIMENT

This section preserves the original design, predictions and numerical results. A training cell denotes one architecture–HUD condition; rendered and hidden denote whether the HUD is visible during training. Its imagined detector endpoints motivate the simulator-scored experiments in the main text; they do not independently establish semantic erasure or compensatory use.

A.1 ENVIRONMENTS AND AGENTS

Rendered cell. Craftax-Classic (Matthews et al., 2024), the JAX reimplementation of Crafter, rendered at $64 \times 64 \times 3$ with the inventory (HUD) drawn in the lower strip of the frame; 22 achievement signals fire at known moments and define the tech tree `collect_wood`→{`place_table`, `make_wood_pickaxe`, `make_wood_sword`}→ .... **Hidden cell.** The POBAX no-inventory variant (Tao et al., 2025): the whole HUD strip is zeroed, which removes the vitals (health, food, drink, energy) as well as the inventory; the world view, including trees and placed wood, is unchanged. (An earlier draft stated that world wood was also occluded; that option exists only as an unimplemented flag and was never used.)

Pixel-loop arm. Δ-IRIS retrained on Craftax from scratch (transformer L=3, d=512, 8 heads; $\approx$25M parameters), 1000 epochs $\approx 10^7$ environment steps. **Latent-loop arm.** DreamerV3 rebuilt from the reference PyTorch implementation with dyn_deter $= 1024$, dyn_hidden $= 512$, cnn_depth $= 48$, two-layer heads and train_ratio $= 128$ (29.0M world-model parameters), trained to 10^7 environment steps. The config was shrunk from the DreamerV3-XL default (181.6M parameters) after the first gate revealed a confound (§A.4); the shrunk model is also the better capacity match to Δ-IRIS. Per cell we pre-registered 5 seeds; after the 5-seed per-cell tests came out non-significant we extended the DreamerV3 cells to 12 (rendered) and 8 (hidden) seeds and the Δ-IRIS cells to 10 (rendered) and 7 (hidden), the counts a power analysis at the observed effect size called for. The extension was decided after seeing those results, so the extended per-cell tests are sequential rather than untouched pre-registered confirmations; both sample sizes are reported. In all, 57 agents were trained (37 in the reported arm, 20 in matched-budget controls, App. C).

A.2 Three measurement tiers and architecture-specific detectors

Decode. Episode-disjoint linear probes on 47 targets (action, reward, and just-unlocked and cumulative achievements), evaluated by AUROC at every tap: for Δ-IRIS the raw codes, the frame embedding, the latent embedding and each transformer block; for DreamerV3 z_t, h_t, their concatenation (what the actor reads) and the CNN embedding of the current frame. **Write.** Projection ablation $h' = h - (h \cdot \hat{d})\hat{d}$ of a concept direction at a tap, scored by the change in log-probability of the codes (Δ-IRIS) or of the realised next latent z_{t+1} under the prior (DreamerV3) that the model actually produced, at the ablated step (gap 0) and three steps later (gap 3). **Drive.** The behavioural cascade: imagination is burnt in on real history ending at a `collect_wood` unlock (burn-in 8) and rolled out closed-loop for 12 steps with the lesion sustained; the outcome is the probability that the wooden-tool-tier detector fires within the horizon, and cascade $=$ baseline $-$ lesioned (positive means the lesion suppressed the tier). In DreamerV3 the lesion is applied at z_t only, h_t only, or both. Each moment's outcome is the mean of four imagination rollouts, each seeded by (moment, site, lesion, replicate), so that adding conditions to a run cannot re-draw the noise of the others (App. E).

Detectors. The tier detector is the object the cascade reads out, so its false-positive rate must be measured, not assumed. For Δ-IRIS it is the top-2 mutual-information codebook atoms per achievement, whose measured random-atom null is 0.020. That detector did *not* transfer to DreamerV3's 32×32 latent, where random atoms fire 32% of the time and 9/10 cells would have "passed" while only 1/10 beat their own null; for DreamerV3, G1 therefore uses a linear probe calibrated to a horizon-level false-positive rate of 0.05 on clean held-out windows (held-out AUROC 0.94–1.00). The cascade refits that probe on each seed's SELECTION corpus with a *per-state* threshold (the 95th percentile of scores on negative states); its horizon-level false-positive rate in the pre-tier states where cascades are evaluated is far higher and is measured in §5.

A.3 Gates, the pre-registered experiment, and the matched control

G0. The *Crafter experiment* denotes the earlier application of this analysis pipeline to a Δ-IRIS agent on Crafter, whose seed-0 checkpoint and fourteen pre-registered targets App. B lists; the unmodified pipeline was first re-run on that checkpoint and reproduced all 14 hard targets within tolerance. **G1.** For every (agent, environment, seed) we measure the baseline closed-loop imagined wooden-tool rate; cells below the pre-registered threshold 0.50 are excluded from the cascade a priori and reported as "no behaviour to break", never silently dropped.

The five predictions (Table 3) were frozen under tag `prereg-v1` together with the selection rules: TRAIN/SELECTION/TEST episode-disjoint; the `collect_wood` direction selected *two ways*, by decodability (detector lift, or the probe) and by causal effect on held-out SELECTION episodes—teacher-forced imagination for Δ-IRIS, closed-loop actor-sampled imagination for DreamerV3 (the same rollout type as TEST, on disjoint episodes)—with the selection rank of the chosen gate reported; imagined return never interpreted under the entangled lesion; and the circularity of selection statistic and outcome flagged throughout.

Variance-matched null. For every causal contrast we build, in addition to random directions, a set of unit directions whose projection variance on the held-out activation corpus lies within $\pm 10\%$ of the SAE direction's ($\pm 5\%$ at fit time), whose cosine with it is at most 0.2, built on one half of the

Table 3: The pre-registered predictions and observed contrasts. The source column records the original hypothesis, not a validated information-erasure result. Cascade = baseline − lesioned tool-tier rate under a sustained lesion of the selected `collect_wood` direction, closed-loop. §5 shows that this direction carries the collection *event* (probe gates) or little wood information at all (dictionary gates), not wood availability. Per-cell tests are paired Wilcoxon signed-rank against the random-direction condition over G1-passing seeds; "own null" is Fisher's combination over the cell's seeds of each seed's empirical p, the rank of its gate among up to 50 variance-matched directions built against that seed's own gate.

	cell	hypothesised source	predicted	seeds (pass/trained)	verdict
P1	Δ-IRIS · rendered	the frame	no cascade	3/5 (prereg), 4/10 (all seeds)	✓ $+0.005$, $p = 0.75$
P2	Δ-IRIS · hidden	nowhere	cascade	0/7	excluded a priori (G1)
P3	DreamerV3 · rendered, z_t only	h_t (hypothesis)	no cascade	10/12	✓ $+0.005$, $p = 0.11$ (1/17 of joint)
P4	DreamerV3 · rendered, $h_t{+}z_t$	none (hypothesis)	cascade	10/12	✓ $+0.077$, $p = 0.002$; own null $p = 1 \times 10^{-4}$
P5	DreamerV3 · hidden, $h_t{+}z_t$	h_t (forced)	cascade	7/8	✓ $+0.128$, $p = 0.016$; own null $p = 8 \times 10^{-6}$

corpus and verified on the other. Every accepted direction (up to 50) is rolled out; a seed's evidence that its direction is privileged over other high-variance directions is the rank of its effect among the null draws, an empirical p, and seeds are combined by Fisher's method. The module also records a *feasibility ratio* $\lambda_\perp / s^2$ (top eigenvalue of the covariance deflated by the feature, over the feature's projection variance): when it is $\lesssim 1 - \text{tol} - \cos^2_{\max}$ no matched distractor can exist, and the test is reported as not runnable rather than as a pass (App. D).

A.4 Imagined competence and the architecture comparison

A confound found and removed. On the first checkpoints only Δ-IRIS·rendered passed G1, but Δ-IRIS had seen ten times DreamerV3's environment steps (at a matched 10^6 budget: 0.297 vs. 0.302, $p = 0.95$), so DreamerV3 was shrunk, which also improved the capacity match, and both arms were trained to 10^7 steps (App. C).

The interaction. At matched budget and capacity (App. C, Fig. 4), hiding the inventory costs the pixel-loop agent -0.350 of its imagined tool-tier rate ($0.568 \to 0.218$, 62%, Welch $p = 0.008$) and the latent-loop agent -0.000 ($0.553 \to 0.553$). The arm×cell interaction is $+0.350$ (OLS $p = 0.005$; permutation $p = 0.017$ on the pre-registered 5 seeds per cell). Over all 37 agents (Δ-IRIS 10 rendered, 7 hidden; DreamerV3 12, 8) the pixel-loop drop is -0.314 ($0.507 \to 0.193$, again 62%, $p = 0.001$), the latent-loop change $+0.004$ ($p = 0.89$), and the interaction $+0.318$ (OLS $p = 0.0008$, permutation $p = 0.004$). Each arm's drop uses its own detector, so the interaction is a difference-in-differences that cancels a detector family's bias only if the bias is additive and equal in both cells of the arm—an untested assumption.

Evidence consistent with an interface account. The decode tier (Fig. 5, §A.5) shows that the HUD mask removes a similar amount of decodable information from the pixels in the two arms: the frame-only representation (Δ-IRIS frame embedding, DreamerV3 CNN embedding) loses 0.111 AUROC in each (difference 0.000, 95% CI $[-0.032, +0.032]$, five agents per cell). What reaches the policy diverges: Δ-IRIS's actor acts on decoded frames and its input representation loses 0.111 AUROC; DreamerV3's actor reads $[z_t, h_t]$, dominated by h_t, and that input representation loses 0.038 AUROC ($p = 0.002$). The readout and behavioural patterns are consistent, and the frame-only control argues against unequal loss of decodability as the explanation. It does not remove the architectural, training-objective or detector confounds.

The gate is unstable at threshold for Δ-IRIS: on a protocol-identical re-draw two seeds cross 0.50 in each direction, so every Δ-IRIS verdict is reported under both draws, and P1 holds under each (Table 5, App. L).

A.5 DECODE TIER: THE HUD SHORTCUT REPLICATES AND RELOCATES

In Δ-IRIS on the rendered cell (Fig. 5) the frame embedding alone (AUROC 0.950) out-decodes every transformer block (0.938–0.942): the current frame suffices and the tested block readouts do not improve on it, replicating the "HUD shortcut" from the Crafter experiment in a new environment. On the hidden cell the frame embedding collapses hardest (-0.109) and drops *below* every block (0.841 vs. 0.867–0.875): the shortcut is gone. Frame-over-blocks dominance falls from 42.5% to 10.1% of (label, seed) pairs, a $4\times$ attenuation.[1] In DreamerV3, h_t (0.939) far out-decodes z_t (0.727) in both cells, the concatenation the actor reads tracks h_t (0.935), and hiding the inventory moves them by only 0.04 (both taps are 1024-dimensional).

Dictionary tier. DreamerV3's z_t is exactly 32-sparse, so its dictionaries use $k = 32$, not the pre-registered 16 (App. F).

A.6 WRITE TIER: WHAT THE VARIANCE-MATCHED CONTROL REMOVES, AND WHAT SURVIVES

Δ-IRIS. The headline single-step ablation from the Crafter experiment replicates: the largest concept effect at gap 0 is again `wake_up`, $+5.31$ nats (feature minus random), against $+4.85$ on Crafter and inside the pre-registered $\pm 20\%$ band, and the read–write asymmetry reproduces, with $+0.45$ nats at gap 0 retained at $+0.3\%$ by gap 3 in the rendered cell (-0.8% hidden). The effect is concept-specific and bidirectional (only 55% of (concept, seed) pairs have feature $>$ random), so the pooled mean is not a meaningful summary; the unit of claim is the concept.

Under the variance-matched null the picture changes. Against random directions, effects of ± 0.1 nats look causal; against a direction matched on explained variance the 99th-percentile threshold is 0.47–1.13 nats, about $10\times$ larger, and only **17/82** rendered and **7/31** hidden (concept, seed) pairs clear it: **79% of single-step contrasts do not exceed directions matched on explained variance**. What survives is the headline: `wake_up` clears its matched null in 4/5 seeds ($+5.41$), `make_wood_sword` and `make_stone_sword` in 3/5. The cascade focus `collect_wood` does not survive in either cell and its raw write effect is near zero (-0.03). This write statistic is distinct from the cascade outcome in P1. The null accepted fewer than the pre-registered 50 directions (App. D).

DreamerV3: h_t dominates the write; z_t is not inert. Projecting a concept direction out of z_t, h_t or both and scoring the realised z_{t+1} gives raw effects up to $20\times$ larger at h_t, but random directions show a $\sim 30\times$ gap too, because h_t feeds the GRU directly. Normalised by each site's own random baseline, the concept direction is $4.4\times$ ($10.0\times$) above random at z_t and $25.7\times$ ($57.8\times$) at h_t in the rendered (hidden) cell: a direction-specific asymmetry of $5.8\times$ ($p = 0.011$, 0.019), with z_t still significantly above its baseline ($p < 0.001$). P3 is therefore supported comparatively ("z_t contributes six-fold less"), and the write tier *predicts* rather than *entails* P4, which must be settled at the drive tier. The matched null is buildable for one concept only (App. D).

A.7 IMAGINED CASCADE EFFECTS BY INTERVENTION SITE

Every cascade in this section is a change in the imagined detector statistic; §5 measures that statistic on real windows without a tier event and tests the lesion against real behaviour.

P1 (Δ-IRIS, rendered): no cascade. Over the three pre-registered G1-passing seeds the causally selected SAE gate gives a cascade of $+0.015 \pm 0.003$ against $+0.009$ for random (paired difference $+0.005$, $p = 0.75$); CAV $+0.014$; teacher-forced ≤ 0.016 in every seed. Under the verdict that covers all ten seeds with one protocol (the 3000-episode draw; four pass) the gate is $+0.001 \pm 0.014$ vs. -0.003 ($p = 0.63$). Across all ten seeds at 400 test moments each, every G1-passing gate cascade lies in $[-0.012, +0.018]$. Lesioning that direction in the pixel-loop agent does not detectably suppress the imagined tool tier: the original "f187" null replicates on a new environment, and it is the *predicted* outcome if the rendered frame still carries the copy. Two observations are reported rather than pooled: the decodability-selected direction in seed 4 cascades ($+0.195$) where its causal gate does not, and the one large gate effect in the arm ($+0.375$) is in a seed excluded at G1. **Decodability**

[1] The Crafter experiment found this pattern for 16/18 hand-picked concepts; our denominator is all 181 estimable pairs, so the levels are not comparable. The rendered→hidden contrast is the replication.

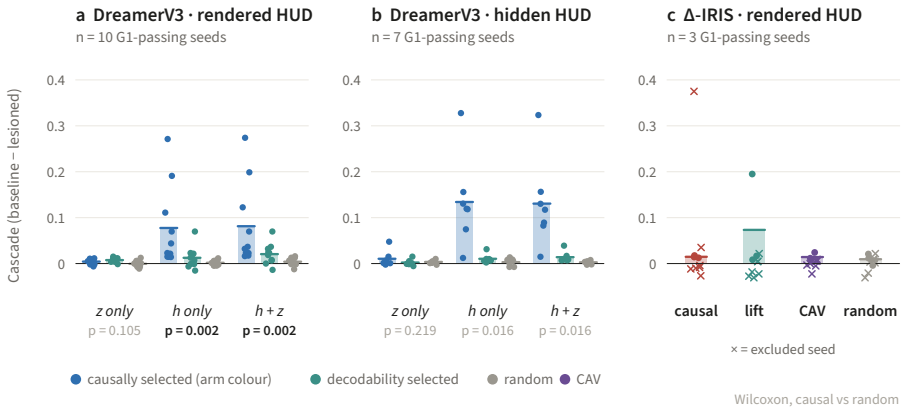


Figure 3: **Imagined cascade effects by architecture and intervention site.** Cascade (baseline − lesioned imagined tool-tier detector rate) per G1-passing seed; the bar is the mean. **(a, b)** DreamerV3: lesioning the causally selected direction in h_t, or in $h_t{+}z_t$, moves the detector in both cells; z_t alone, the decodability-selected direction and random do not. p is the paired Wilcoxon test, causal vs. random. **(c)** Δ-IRIS with the frame still rendered (P1), closed-loop, 400 test moments per seed: no lesion separates from random among the G1-passing seeds (○); the one large gate effect ($+0.375$) is in a seed excluded at G1 (×). G1 verdict from the recorded rollouts: 3/9 seeds pass.

$\neq$ **generative control** replicates in 10/10 seeds: the causally strongest feature sits at lift rank 2–12, never 1.

P2 (Δ-IRIS, hidden): excluded. All seven seeds fail G1 (0.05–0.39): there is no tool-making behaviour to break, mirroring the no-HUD exclusion in the original audit. The within-Δ-IRIS contrast is therefore untestable, and the discriminative burden falls on the latent-loop arm.

P3 (DreamerV3, z_t only): almost no cascade. Over the ten passing rendered seeds the z_t-only gate lesion gives $+0.004 \pm 0.006$ against -0.001 for random (difference $+0.005$, Wilcoxon $p = 0.11$; Fig. 3); in the hidden cell $+0.010$ vs. $+0.004$ ($p = 0.22$). Pooled over 17 seeds the effect is small but detectable ($+0.006$, $p = 0.029$), one-seventeenth of the joint lesion's. P3 holds in its comparative form—h_t-only exceeds z_t-only by $+0.094$ ($p = 1.5 \times 10^{-5}$)—and a direction that decodes from z_t at AUROC 0.73 moves the imagined tier an order of magnitude less when only z_t is lesioned.

P4 and P5 (DreamerV3, $h_t{+}z_t$): cascade. The joint lesion of the causally selected gate suppresses the detected tier in the rendered cell ($+0.081 \pm 0.090$ vs. random $+0.004$; paired $+0.077$, Wilcoxon $p = 0.0020$, t-test $p = 0.027$, $d = 0.83$, gate $>$ random in 10/10 seeds) and in the hidden cell ($+0.131 \pm 0.096$ vs. $+0.003$; $+0.128$, $p = 0.016$, t-test $p = 0.012$, $d = 1.34$, 7/7). Pooled ($n = 17$): $+0.098$, $p = 1.5 \times 10^{-5}$, $d = 1.04$. At the five seeds per cell originally trained neither per-cell test was significant ($p = 0.17$, 0.11), and with one rollout per moment the rendered test moved from $p = 0.004$ to 0.049 when an unrelated change re-drew the noise; the values above use the power-analysis seed count and four rollouts per moment, and every draw is recorded (App. E).

The joint and recurrent-only effects are close. Lesioning h_t alone gives $+0.099$ over random ($p = 1.5 \times 10^{-5}$ pooled) and adding z_t changes it by $+0.001$ ($p = 0.20$); h_t-only exceeds z_t-only by $+0.094$ ($p = 1.5 \times 10^{-5}$). The "joint" condition's effect is not distinguishable from the recurrent-site lesion's, and the recurrent site is where §A.5–A.6 located the state.

Selection and the matched null. The decodability-selected direction cascades six-fold less than the causally selected gate ($+0.018$ vs. $+0.102$, $p = 0.0005$) and is the causally strongest in only 4 of 17 seeds. Against a variance-matched null built for each seed's own gate, the gate beats every draw in 9/10 rendered and 7/7 hidden seeds (Fisher $p = 1.4 \times 10^{-4}$, 8.3×10^{-6}; App. D, E).

A.8 A LESION'S FOOTPRINT ON THE NEXT STATE PREDICTS THE CASCADE; RESTORATION REACHES ONE STEP BACK

Disturbance and cascade sensitivity. The pre-registered link between the cells is a re-derivation residual: how far a lesion moves the next state the policy reads—the mean absolute change of h_{t+1} (DreamerV3) or of the decoded HUD strip (Δ-IRIS) on paired rollouts that are identical without a lesion, using each cascade's own directions and TEST moments. In DreamerV3 it predicts the cascade across 153 lesions in 17 seeds (App. H; Spearman $+0.53$, seed-clustered 95% CI $+0.45$ to $+0.61$) and within the causally selected gates alone ($+0.70$), so it is not explained solely by separating gate and random directions. In Δ-IRIS it points the same way but cannot be tested (four seeds pass G1 and one lesion cascades; $+0.42$, CI -0.14 to $+0.66$), and no lesion reaches the pre-registered $3/255$ HUD threshold. Across agents the cascade covaries with a lesion's whole-state footprint and weakly with the measured wood-readout losses (F5, App. K), so we read the residual as a disturbance measure, not as evidence that the lesion removed the copy.

Prediction-restoration horizon. Causal tracing restores clean information at one step of a corrupted window and measures recovery of the next-latent prediction (App. H). Δ-IRIS is quasi-Markovian on rendered Craftax (0.43 from the current step, ≈ 0.03 from any earlier one). DreamerV3 restores $2.5\times$ more one step back (0.111 vs. 0.045, Holm $p = 0.0001$) but is not distinguishable from Δ-IRIS two steps back or beyond: on this next-latent restoration task its recurrent state extends causal reach by one step, not the long reach we predicted—a statement about one task, not about memory capacity. Hiding the inventory does not raise Δ-IRIS's restoration from history (0.028 vs. 0.045), so the original's no-HUD increase does not replicate.

B G0: REPRODUCTION OF THE ORIGINAL AUDIT

Before anything was changed, the unmodified analysis pipeline was run on the original Crafter seed-0 checkpoint through the refactored harness and compared to fourteen pre-registered targets with tolerances: best-detector lift (707.5 vs. 708), per-slot active codes (exact), SAE FVU (0.044 vs. 0.043), HUD shortcut ($16/18$), maximal ablation ($+4.85$ nats on `wake_up`), empirical-null z (24.2), the honest negative (`collect_iron`, -2.3), direct logit attribution ($0.311/0.083/0.013$), the f187 cascade null (wood $1.00 \to 0.03$, tools $0.80 \to 0.78$), and causal-tracing restoration (0.001 history, 0.495 final action token). All 14 hard rows and 3 soft rows passed.

C THE TRAINING-BUDGET CONFOUND

Δ-IRIS's collection schedule yields $\approx 10^7$ environment steps at 1000 epochs, while the plan's DreamerV3 target was 10^6. Matching Δ-IRIS *down* to 10^6 (91 epochs; 10 additional agents) gave rendered 0.297 vs. DreamerV3-XL 0.302 ($p = 0.95$, $d = -0.04$); within Δ-IRIS the budget effect was $0.297 \to 0.568$ ($p = 0.036$, $d = 1.60$). At 10^6 only 1/20 cells passes G1, so a matched-budget study at that budget has nothing to break. Training the XL DreamerV3 to 10^7 was measured at ≈ 46 days per agent; the cost is gradient-step count, not FLOPs, so the configuration was shrunk (world model 181.6M$\to$ 29.0M; train_ratio $512 \to 128$; ≈ 5.5 days per agent). At equal gradient steps the shrunk model outperforms the XL one at every checkpoint measured (5.36 vs. 4.82 return at 125k steps, rendered).

D THE VARIANCE-MATCHED NULL: WHEN IT CAN AND CANNOT BE BUILT

The worry that "matching variance forces similarity" conflates a subspace property (variance height) with a ray property (identity). The constraints conflict only when the variance regime at the feature's height is effectively one-dimensional, i.e. when $\lambda_\perp / s^2 \lesssim 1 - \text{tol} - \cos^2_{\max}$. A synthetic sweep with the feature inside a top-variance subspace of dimension k gives feasibility ratios $0.034, 0.986, 1.05, 1.06, 1.17$ and acceptance $0/30, 12/30, 30/30, 30/30, 30/30$ at $k = 1, 2, 4, 16, 64$, with all accepted directions exactly orthogonal to the feature. Δ-IRIS's residual stream is the k-large regime; DreamerV3's taps are often $k \approx 1$, where the hypotheses "the identity matters" and "being the top-variance direction matters" coincide and no control can separate them. A separate sampler-reach failure (ratio $\gg 1$, the feature far below the bulk variance) occurred for

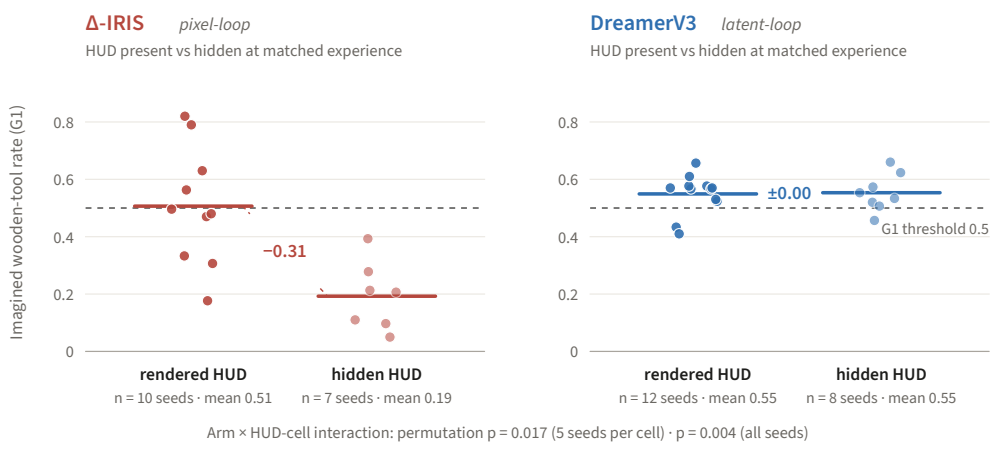


Figure 4: **Hiding the HUD has different effects on the two agents' imagined competence.** Imagined wooden-tool rate (G1) for every seed; the horizontal bar is the cell mean, the dashed line the pre-registered G1 threshold. Δ-IRIS drops by 0.31 over all seeds (0.35 at the pre-registered five per cell) when the inventory HUD is hidden; the estimated DreamerV3 difference is small. Arm×cell interaction by permutation: $p = 0.017$ at the pre-registered 5 seeds per cell, $p = 0.004$ with all 37 agents. Δ-IRIS seeds 0–4 are shown at their pre-registered draw, later seeds at their only draw.

the Δ-IRIS cascade gate (1/50 accepted, ratio ≈ 43); there the cascade comparison is not needed because the gate is already indistinguishable from random. For the DreamerV3 write tier the matched null could be built for one concept ($p_{99} = 0.09$ against an effect of 1.94); for the others the feasibility ratio says it cannot exist. The diversity of accepted directions is also recorded, since at small k they can collapse onto one ray (pairwise $|\cos| = 0.97$ at $k = 2$). In the Δ-IRIS write tier the null accepted $27.0/50$ (rendered) and $20.6/50$ (hidden) directions on average after both constraints, and two concepts admit none; constraints were not relaxed to reach 50.

E Measurement errors caught by the gates, and their fixes

Each item below changed a number that would otherwise have been reported. (1) The MI-atom tier detector did not transfer from a 4×1024 codebook to a 32×32 latent (random-atom null 0.74–0.94); replaced by an FPR-calibrated probe for DreamerV3. (2) The ≥ 8-step burn-in filter anti-correlated with competence (competent agents reach wood by step 3), leaving as few as 3 moments per seed from 150 episodes; fixed by 3000-episode corpora, $n = 100$ everywhere. (3) At 3000 episodes the gate was OOM-killed on three cells and the aggregation silently read stale outputs; fixed by decoupling episodes encoded from episodes scanned, and by deleting outputs before re-running so a crash leaves a missing file, not a plausible old one. (4) SAE $k = 16$ was a spec mismatch for a 32-sparse latent (App. F). (5) Hidden-cell SAEs trained on $4\times$ less data, a confound in the hypothesis's favour; fixed to sample parity. (6) The raw h_t/z_t write asymmetry conflated site and direction; fixed with a per-site random baseline, which overturned an earlier "z_t is write-inert" reading. (7) The Δ-IRIS teacher-forced outcome was a tautology in its first form (always one); replaced by a probabilistic outcome. (8) The DreamerV3 cascade gate initially had no matched null of its own; one was added. (9) The Δ-IRIS arm's cascade first ran at 60 test moments (n≈48 available) and its new seeds were gated on 150-episode corpora; both were re-run at the DreamerV3 arm's settings (3000 episodes, 400 test and 120 selection moments), which is the analysis reported here. (10) A scratch auto-delete removed the Δ-IRIS seed 0–4 corpora and gate files; the recorded verdicts are carried in the repository and the corpora were regenerated (§A.4). (11) The cascade's matched null rolled out only 10 of up to 50 accepted directions, so its "99th percentile" was the maximum of 10 and "exceeds p99" a 1-in-11 rank event per seed; every accepted direction is now rolled out, the full null recorded, and seeds combined by Fisher's method. A pairing bug in the joint-site null (draw j of z_t with draw j of h_t, which accepts fewer, so the tail became z_t-only lesions)

is fixed. (12) Re-running with more nulls re-drew the imagination noise of every later condition, because all rollouts shared one RNG stream: P4 read $p = 0.004$, then 0.049, on identical moments. Each moment's outcome is now the mean of four rollouts, each seeded by (moment, site, lesion, replicate); the third run ($p = 0.002$) is reported and all three are kept. (13) Which direction the causal selection returns changed between random-number streams in some seeds. Re-ranking the same seven candidates on the same selection moments under ten fresh rollout streams (at one and four rollouts per moment), recording their cosines, and rolling every candidate out on TEST shows why: the top-1/top-2 margin in units of its own standard deviation predicts how often the modal gate wins (Spearman $+0.91$, $p < 0.001$, 17 seeds). Six seeds (median margin/SD 5.1) select the same gate on all ten streams, with a large TEST cascade ($+0.190$); in the eleven unstable seeds (median 0.5) winner and runner-up are near-tied on TEST ($+0.054$ vs. $+0.042$, random $+0.005$) and near-orthogonal (median $|\cos| = 0.04$). Four rollouts per moment raise modal agreement from 79% to 89%. The selection statistic ranks all candidates as TEST does (median within-seed Spearman $+0.89$), and the reported gate matches the modal gate in 15/17 seeds; the two exceptions are unstable seeds whose cascades are of the same size either way. (14) The Markovianity summary first took the maximum restoration over token groups and blocks; on the smoke test this gave Δ-IRIS a flat floor of ≈ 0.07 at every distance—the maximum of nine noisy cells against DreamerV3's three—enough to make a quasi-Markovian curve look history-dependent. The reported estimator is the mean over cells, fixed on one seed per arm before any full run, with the maximum curve kept as a sensitivity check; the five Δ-IRIS hidden corpora lost to the purge were re-collected for tracing.

F Dictionary tier: the pre-registered sparsity does not transfer

TopK SAEs ($M = 2048$, 80 epochs) at $k = 16$ reach FVU 0.031 on Δ-IRIS's block-1 residual stream (better than the original 0.043) but 0.20–0.32 on DreamerV3's taps. The cause is structural: z_t is exactly 32-sparse by construction, so a dictionary with $k = 16$ has fewer active atoms than the target has active dimensions. A sweep shows a $3\times$ cliff for z_t precisely at $k = 32$ ($0.181 \to 0.060$) and no cliff for the continuous h_t; doubling M makes z_t worse and creates dead features, so sparsity, not capacity, binds. DreamerV3 dictionaries therefore use $k = 32$ (FVU 0.033–0.035 for z_t, 0/2048 dead), a recorded deviation. A second issue was caught here: hidden-cell dictionaries first looked worse (0.213 vs. 0.035) but had trained on $4\times$ less data because only rendered corpora had been enlarged; re-collecting to sample parity closed the gap (0.0334 vs. 0.0354). Left standing, that confound would have biased every hidden-cell contrast in the direction of the hypothesis (App. E).

G Decode map

Figure 5 reports held-out inventory readouts across the two architectures. The source-specific gaps motivate the crossed interventions in the main text; they do not establish which information the policy uses.

H State disturbance and prediction-restoration horizon

The panels are described in App. A.8. The cascade in (b) and (c) is the imagined detector's; App. K (F5) shows that across agents it tracks the size of a lesion's disturbance rather than the concept information it removes, so (b) is reported as a disturbance–sensitivity relation.

I Phase 7: the follow-up protocol in full

Why a second protocol. The pre-registered architecture experiment uses a calibrated detector reading an imagined latent as its endpoint, and its lesion direction was selected by causal effect on an imagined outcome. Three questions are therefore open: what information the intervention removes, whether its effect exists outside the detector, and whether a surviving copy explains the resistance to a z_t-only lesion. The protocol (`runs/phase7/PROTOCOL_2026-09-17.md`) was written and dated before the corpora it uses were collected. It separates confirmatory from exploratory contrasts, freezes eligibility (the 17 G1-passing DreamerV3 agents and the 4 Δ-IRIS agents of the Phase-6 regression), and fixes an engineering pilot on the two lowest-numbered eligible seeds per cell. Pilots

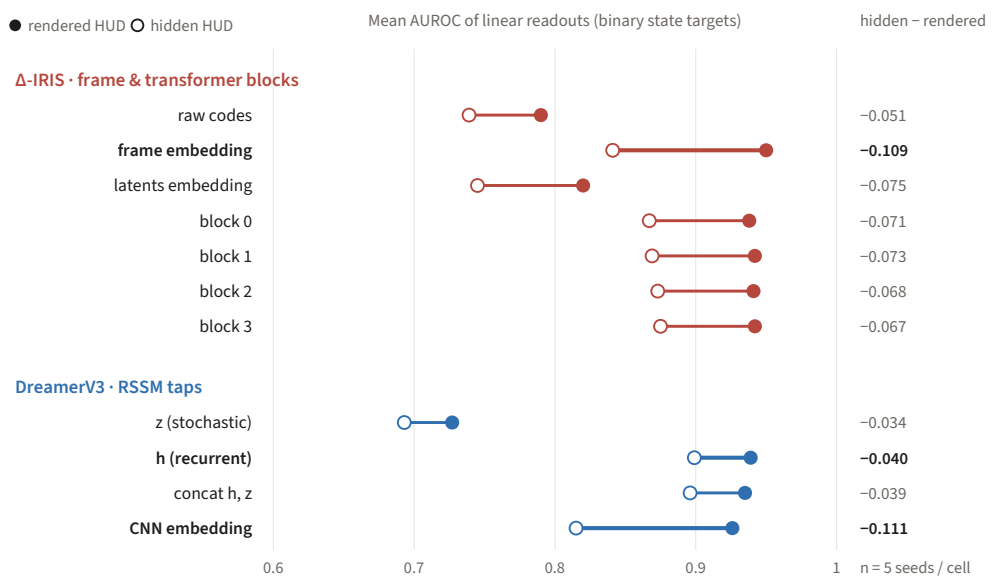


Figure 5: **Inventory information is readable at different sites in the two agents.** Decode map: mean AUROC over binary targets, 5 seeds pooled per cell. Filled dot = rendered Craftax, open dot = hidden-state Craftax; the connector length is the AUROC lost when the HUD is hidden. In Δ-IRIS the frame embedding decodes best and loses most (-0.109), falling below every transformer block; in DreamerV3 the recurrent state h_t decodes best and barely moves (-0.040), while the CNN embedding of the frame loses 0.111.

ran on a separate development corpus; its episodes are never used for confirmatory evaluation, but the pilot agents belong to the confirmatory set, whose moments come from a fresh corpus collected afterwards. Later experiments (F2R, F7, and the reruns after the audit of App. K) each have their own dated protocol; Table 8 lists every run with its status, and only primary runs support claims.

Corpora and pairing. Each agent has a development corpus, used only by the pilots, and a confirmatory corpus collected afterwards; the two are never mixed. Within the confirmatory corpus, readouts, subspaces, donors and thresholds are fit on even episodes (the design split) and every evaluated moment comes from odd episodes. Each records the agent's own recurrent state and sampled code, the full inventory, the game's own craftability conditions read from `game_logic.py` (a wooden pickaxe or sword needs wood$\geq$1 and a nearby table; a table needs wood$\geq$2; placing stone needs stone$\geq$1 and a free faced cell, recorded conservatively), the episode's PRNG key and a hash of every state leaf. Replaying the key and the recorded actions re-creates any frame's simulator state exactly, verified against the stored hash, so a real-environment branch starts from the state the agent actually saw. Moments are first wood (or stone) collections in evaluation episodes. Conditions within a moment share the branch, the exogenous randomness and the torch seed, so they differ only by the intervention; outputs record the moment ids and their hash.

Consumers (F1). The lesion is applied to the state that one consumer reads—the detector, the actor, or the transition—or to all of them, in one-shot and sustained form. In the real environment the recurrent state is threaded through the chosen consumer at every step and the simulator scores achievements, tools, wood spent and death; there is no detector condition there by construction, and no imagined effect is subtracted from a real one.

Availability subspace and donors (F2, F3R2, F7). W_{h_t} is a rank-revealing basis of the raw-space directions of logistic availability readouts fit on the design split: {wood$\geq$1, wood$\geq$2, craftable wooden pickaxe, placeable table} for wood and {stone$\geq$1, craftable stone pickaxe, placeable stone} for stone (the stone sword label duplicates the pickaxe label and is dropped). Each standardised-space readout coefficient is divided by its activation standard deviation to obtain a raw-state direction and an SVD keeps singular values above 10^{-2} of the largest; the rank is 4 in every agent

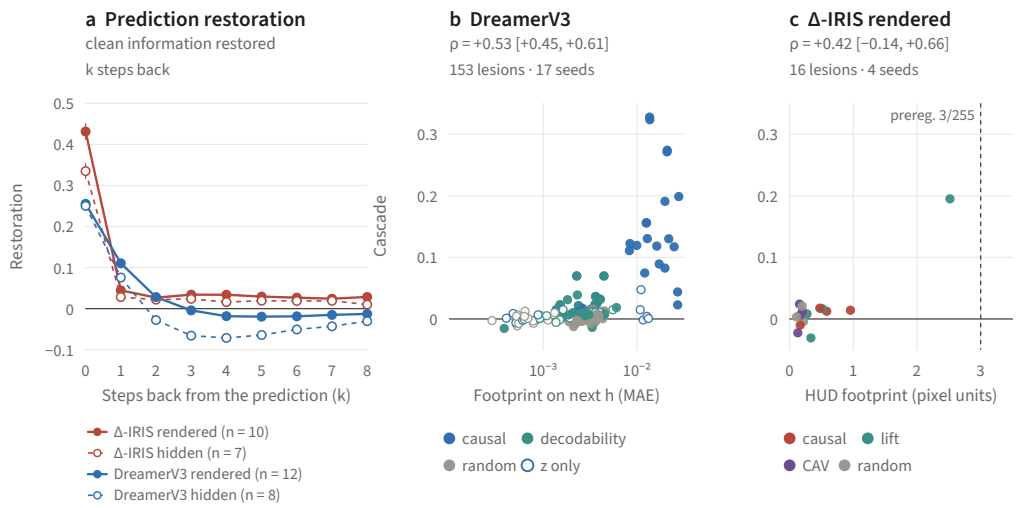


Figure 6: **The imagined cascade grows with how far a lesion moves what the policy reads next.** **(a)** Causal tracing: restoration of the next-latent prediction by clean information k steps back (mean over token groups, ±SEM over agents). Δ-IRIS reads essentially only the current step; DreamerV3's recurrent state carries one step more. **(b)** DreamerV3: cascade against each lesion's one-step footprint on h_{t+1}, all lesions in G1-passing seeds; open dots are z_t-only lesions. **(c)** Δ-IRIS: cascade against each lesion's footprint on the decoded HUD strip; the dashed line is the pre-registered 3/255 threshold. ρ is Spearman's correlation with a 95% interval resampling seeds.

for wood and 3 for stone. F2's unrelated control U_{h_t} is the same construction over stone, coal, sapling and stone-pickaxe labels; it had rank 3 in 12 of 17 agents, so that control is not rank-matched (a later code change rank-matches it; the historical control design was reproduced, not replaced by a rank-matched design). F3R2's donors are design-split frames matched to the receiver on the concept's availability, craftability and placeability flags, table proximity and three non-target flags (the other resource's availability and tool possession), with a fallback recorded per step; the wrong-status donor flips the availability flag and the flags that depend on it. As recorded flags, the keys are, for wood, `wood_ge1`, `wood_ge2`, `can_make_wood_pickaxe`, `can_place_table`, `near_table`, `has_wood_pickaxe`, `has_wood_sword`, `stone_ge1`; for stone, `stone_ge1`, `can_make_stone_pickaxe`, `can_place_stone`, `near_table`, `wood_ge1`, `has_stone_pickaxe`, `has_wood_pickaxe`. `facing_wall` is in neither key (§2). At z_t a projection leaves the one-hot support and almost never changes the sampled code, so the z_t manipulation overwrites the eight slots with the largest availability weight with a donor's valid codes.

Margin and horizons. "No extra effect" is tested against a ± 0.05 margin on the tier rate, set from the pilots' interaction spread (sd ≈ 0.03) before the confirmatory corpora existed, and equivalence is claimed only when the whole 90% interval lies inside it. The real-environment horizon for F1, F2, F4 and F7 was fixed by a pre-stated rule (the smallest of 25 and 50 steps at which the pooled intact unlock rate on the pilots exceeds 0.8; it gave 25, the pooled rate being 0.849 against 0.487 at 13 steps and 0.993 at 100, where the endpoint saturates). The repair experiments score 13 steps, fixed with F3's contrasts, where the intact rate leaves room for effects in both directions.

Mechanical deviations this protocol found in the pre-registered architecture experiment. (i) The dictionary lesions project out $\mathrm{unit}(w)$ for a decoder column w, but the sparse autoencoder is fit on standardised activations, so the feature's raw displacement is $\mathrm{unit}(\sigma \odot w)$; the cosine between them is 0.92–1.00 at z_t and 0.82–0.98 at h_t. Every Phase-7 lesion uses the standardised convention, and the gate from that experiment is carried as a fixed comparator. (ii) The z_t-site projection leaves the simplex: about half the entries go negative while the per-slot argmax changes in only 0.01–0.25 of 32 slots, and after re-one-hot the probe gate's z_t-site readout effect vanishes. (iii) Δ-IRIS's world model is not in its test-time control loop, so a world-model lesion cannot change its real behaviour;

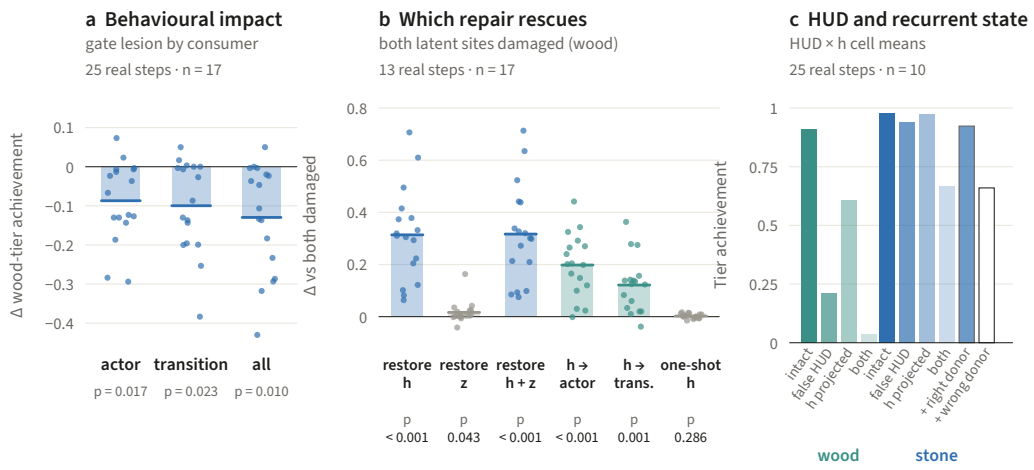


Figure 7: **Interventions scored by the simulator** (DreamerV3; dots are agents; p Holm-corrected). **(a)** The gate lesion routed to the actor only, the transition only, or both lowers real wood-tier achievement within 25 steps. **(b)** After damaging the recurrent availability subspace and the code's wood slots: repairing h_t rescues, through the actor and through the transition; repairing z_t has a small effect whose failed positive control prevents a mechanistic conclusion; a single first-step repair under continuing damage does not rescue (13 steps). **(c)** Rendered agents, bars left to right: intact, a false zero count on the HUD, the recurrent subspace removed, both; for stone, two more with the count falsified (outlined): the subspace taken from a correct-status donor, and from a wrong-status donor. For stone the joint impairment is disproportionately large and the correct-status subspace largely reverses it; for wood each manipulation impairs on its own.

the pixel-loop cascade is an imagined-representation result by construction, and the behavioural test for that arm is the HUD edit.

J Concept replication and horizon dependence (F6)

The same manipulations with *stone* in place of wood, on the same agents and corpora (protocol addendum dated 2026-09-17, written before the runs). The stone tier is dominated by `place_stone`, which needs stone and a free faced cell but no wood or table; stone tools also need wood and a table and are rare (sham rates: pickaxe 0.123, sword 0.047, placing stone 0.973, over rendered agents).

An implementation error, and what it withdrew. The first stone runs orthonormalised a label set in which two labels are one predicate (a stone pickaxe and a stone sword have the same preconditions) without a rank check, so every stone basis carried an arbitrary direction. Results that depend on that basis—a stone $z_t \times h_t$ interaction of -0.064 in earlier drafts, and the first stone HUD$\times h_t$ run—are withdrawn; the rank-revealing basis (rank 3) replaced it in F3R2 and in a full rerun of F4's DreamerV3 stone conditions. Wood bases were full rank (singular values 1.30–0.49).

Recurrent subspace (F3R2, 13 steps). Stone-tier achievement is near ceiling at 13 steps (0.971 intact) and effects are small: damaging the recurrent subspace $+0.001$, repairing it after damaging both $+0.002$ (Holm $p = 1.0$), so the positive control is uninformative rather than failed; the $z_t \times h_t$ interaction is -0.007 (90% CI $[-0.013, -0.002]$, equivalent to zero within ± 0.05).

HUD (F4, 25 steps). A false zero stone count nearly abolishes stone-tier achievement in Δ-IRIS ($0.989 \to 0.033$; placing stone $0.959 \to 0.022$) and barely moves it in DreamerV3 ($0.979 \to 0.938$, all 10 agents lower, $p = 0.002$); the pooled effect (-0.303) hides this split. In DreamerV3 the joint impairment (0.315) far exceeds the single ones (HUD 0.042, recurrent subspace 0.006; Fig. 7c), which is consistent with partial compensation between the two. Δ-IRIS has no recurrent condition in this design.

Table 4: Primary Phase-7 contrasts, generated from the aggregated artefacts. One number per agent (mean over its evaluation-split moments of the paired difference); means over agents; p: two-sided Wilcoxon signed-rank over agents, Holm-corrected within each family for F1, F2, F2R, F3R2, F4R and F7; unadjusted for the F4 rows (Δ-IRIS, $n = 4$: descriptive, no test); TOST equivalence at ± 0.05 for the $z_t \times h_t$ interaction rows. Endpoint: real concept-tier achievement within 25 or 13 steps scored by the simulator ("tier", mostly `place_table` for wood and `place_stone` for stone), or the imagined detector statistic ("detector"). F4 rows use rendered agents only; F7 uses the 8 rendered agents whose intact fresh actor passed the competence rule. Superseded runs are listed in Table 8, not here.

exp.	contrast	endpoint	n	mean	rendered	hidden	p
F1	actor only − intact	tier, 25	17	-0.087	-0.062	-0.123	0.017
F1	transition only − intact	tier, 25	17	-0.100	-0.072	-0.139	0.023
F1	all consumers − intact	tier, 25	17	-0.130	-0.083	-0.196	0.0096
F1	detector only − intact	detector	17	-0.022	-0.018	-0.027	0.039
F1	actor only − detector only	detector	17	$+0.000$	$+0.008$	-0.011	0.49
F1	transition only − detector only	detector	17	-0.027	-0.015	-0.044	0.039
F2	correct − wrong status (gate lesion)	tier, 25	17	$+0.410$	$+0.413$	$+0.405$	2×10^{-4}
F2	correct status − gate lesion	tier, 25	17	$+0.022$	$+0.030$	$+0.009$	0.0063
F2	correct − unrelated (gate lesion)	tier, 25	17	$+0.006$	$+0.018$	-0.010	0.22
F2R	correct − wrong status (feature lesion)	tier, 25	17	$+0.426$	$+0.438$	$+0.408$	8×10^{-5}
F2R	correct status − feature lesion	tier, 25	17	$+0.026$	$+0.012$	$+0.045$	0.064
F2R	correct − unrelated (feature lesion)	tier, 25	17	$+0.020$	$+0.008$	$+0.036$	0.087
F2R	correct − random (feature lesion)	tier, 25	17	$+0.035$	$+0.016$	$+0.062$	0.0071
F3R2	z_t repair after z_t damage (control)	tier, 13	17	$+0.011$	$+0.014$	$+0.006$	0.10
F3R2	h_t repair after h_t damage (control)	tier, 13	17	$+0.298$	$+0.321$	$+0.265$	0.0011
F3R2	repair h_t after both	tier, 13	17	$+0.314$	$+0.360$	$+0.248$	6×10^{-5}
F3R2	repair z_t after both	tier, 13	17	$+0.016$	$+0.020$	$+0.012$	0.043
F3R2	repair both after both	tier, 13	17	$+0.317$	$+0.367$	$+0.245$	6×10^{-5}
F3R2	repair both − repair h_t	tier, 13	17	$+0.003$	$+0.007$	-0.003	0.46
F3R2	repaired h_t to actor only	tier, 13	17	$+0.198$	$+0.223$	$+0.163$	9×10^{-5}
F3R2	repaired h_t to transition only	tier, 13	17	$+0.121$	$+0.128$	$+0.111$	0.0014
F3R2	one-shot h_t repair	tier, 13	17	$+0.002$	$+0.005$	-0.002	0.29
F3R2	z_t damage	tier, 13	17	-0.070	-0.091	-0.041	0.0011
F3R2	h_t damage	tier, 13	17	-0.281	-0.281	-0.282	0.0011
F3R2	z_t damage given h_t damage	tier, 13	17	-0.076	-0.123	-0.010	0.0021
F3R2	$z_t \times h_t$ interaction (TOST ± 0.05)	tier, 13	17	-0.006	—	—	0.042
F3R2 stone	h_t repair after h_t damage (control)	tier, 13	17	-0.005	$+0.000$	-0.012	0.82
F3R2 stone	repair h_t after both	tier, 13	17	$+0.002$	$+0.009$	-0.008	1.00
F3R2 stone	h_t damage	tier, 13	17	$+0.001$	-0.005	$+0.008$	0.85
F3R2 stone	$z_t \times h_t$ interaction (TOST ± 0.05)	tier, 13	17	-0.007	—	—	6×10^{-11}
F4	HUD false zero − sham, wood, Δ-IRIS	tier, 25	4	-0.918	-0.918	—	—
F4	HUD false zero − sham, wood, DreamerV3	tier, 25	10	-0.707	-0.707	—	0.002
F4	HUD false zero − sham, stone, Δ-IRIS	tier, 25	4	-0.956	-0.956	—	—
F4	HUD false zero − sham, stone, DreamerV3	tier, 25	10	-0.041	-0.041	—	0.002
F4	HUD $\times$ h_t interaction, wood	tier, 25	10	$+0.127$	$+0.127$	—	0.56
F4	HUD $\times$ h_t interaction, stone	tier, 25	10	-0.268	-0.268	—	0.002
F4	stone − wood interaction (paired)	tier, 25	10	-0.394	-0.394	—	0.0098
F4R	restore h_t vs removed, count falsified	stone, 25	10	$+0.258$	$+0.258$	—	0.012
F4R	correct − wrong status, count falsified	stone, 25	10	$+0.263$	$+0.263$	—	0.012
F4R	content effect, falsified − true count	stone, 25	10	$+0.238$	$+0.238$	—	0.012
F4R	F4 interaction replicated	stone, 25	10	-0.268	-0.268	—	0.0078
F7	wood subspace masked − intact	tier, 25	8	-0.049	-0.049	—	0.22
F7	random subspace masked − intact	tier, 25	8	$+0.001$	$+0.001$	—	0.95
F7	wood − random subspace masked	tier, 25	8	-0.050	-0.050	—	0.047

Restoring the stone content while the count is falsified (F4R). A separate dated protocol crossed the HUD edit with four states of the recurrent stone subspace—kept, removed, taken from a correct-status donor, taken from a wrong-status donor (F3R2's donor rule and random stream; one donor of each kind per step in every condition, identical until rollouts diverge)—on F4's moments and seeds; the five conditions shared with F4 reproduce its recorded outcomes exactly. All four primary contrasts hold after Holm correction: with the count falsified, correct-status restoration beats removal ($+0.258$, 9 of 10 agents) and beats wrong-status ($+0.263$, 9 of 10); that content effect is larger than with the true count on screen ($+0.238$, 9 of 10); and F4's interaction replicates (-0.268, 10 of 10). The log-odds versions agree. With the true count on screen the subspace's content makes no detectable difference (right − wrong $+0.026$, right − removed -0.002), and with the count falsified a wrong-status subspace is no worse than none (-0.005). One agent (rendered seed 0, which has the smallest joint impairment) goes the other way. Donor-seed replication (protocol addendum dated

2026-09-20, committed before the runs; donor seed 0 remains the pre-specified primary): at donor seeds 1 and 2 the four primary contrasts keep their signs and their significance after correction—restoration $+0.260$ and $+0.257$ (9 of 10 agents each), correct minus wrong $+0.272$ and $+0.270$ (9 of 10), the display-conditional content effect $+0.248$ (10 of 10) and $+0.245$ (9 of 10), and F4's interaction -0.268 at every seed. Per-agent rank correlations between seeds are $+1.00$ for restoration and $\geq +0.94$ for the content contrast, with mean absolute differences of 0.007–0.027; averaging each agent over the three seeds gives $+0.258$, $+0.269$, $+0.243$ and -0.268. The same agent (rendered seed 0) remains the exception at all three tested donor seeds. All three seeds reproduce F4's five shared conditions exactly. These contrasts support the reading that the recurrent stone subspace carries stone content the policy uses when the on-screen count is wrong, subject to two confounds that remain unresolved: the patches are not matched in size (below), and the donor key omits wall status in front of the agent. Exact correct-status key matching fixes this flag through placeability for receivers with at least one stone or two wood, but not with only one wood. Wrong-status matching provides no such guarantee. Offline agreement is 86% versus 73% for stone in the ten F4R agents and 68% versus 60% for wood over 17 agents (medians over agents). These rates describe sampled donors; they are not a comparison against a validated chance baseline. The compensation is in any case partial, since the false count alone still costs 0.041. Right- and wrong-status donors matched every flag of their keys in approximately 99.98% and 99.99% of cases, respectively (means over agents); fallback was rare, but the patches are not matched in size (0.041 vs. 0.051 of $\|h_t\|$ with the count falsified), so correct $-$ wrong mixes content with a difference in disturbance: the pattern supports a content-based account of partial compensation, but does not isolate resource content from disturbance or collateral effects.

Donor selection targeting matched magnitudes (protocol addendum dated 2026-09-22, committed before the runs). Four added conditions selected a real frame from the same flag-keyed pool to minimise the difference between its patch norm and the paired, opposite-status random donor's norm on that condition's current state. The agents, branching moments and donor seed were unchanged; all nine original conditions reproduced F4R's recorded outcomes in every rollout. Matching failed the pre-specified 10% criterion: median same-step relative errors were 18.5% for wrong-status selection and 10.7% for correct-status selection. The pre-specified behavioural contrasts remain positive: M1 $+0.121$ and M2 $+0.234$ (both Holm $p = 0.041$), and display interactions M3 $+0.119$ ($p = 0.041$) and M4 $+0.234$ ($p = 0.016$), each in 9 of 10 agents. These are effects under imperfect matching, not a passed control.

The matching target is local to each trajectory; it is not the patch actually applied in the comparator's separate rollout. Applied norms are 0.034 (random correct), 0.041 (random wrong), 0.073 (selected wrong), and 0.077 (selected correct), as medians over steps and rollouts within agents, then over agents. Thus M1 compares 0.034 with 0.073, and M2 compares 0.077 with 0.041. The original F4R report's 0.041 and 0.051 instead summarise mean within-rollout displacements before the median over agents; the summaries must not be interchanged. A post-hoc comparison of the two selected conditions gives $+0.092$ (9/10 agents, uncorrected $p = 0.084$); similar aggregate norms (0.077 vs. 0.073) do not imply matched per-step perturbations.

Donor distance varies with intervention history. Along the unpatched false-display trajectory, hypothetical correct- and wrong-status patches have norms 0.079 and 0.046; under sustained projection, their pre-patch distances are 0.088 and 0.043. Along sustained correct restoration they are 0.034 and 0.095, and along wrong restoration 0.100 and 0.041. This challenges an unconditional account in which correct donors are always closer, but does not control disturbance in the restoration comparisons or establish semantic drift. Selected wrong donors give higher achievement than random wrong donors (0.802 vs. 0.660). The independent readout validation samples random donors, so the target and collateral effects of the selected donors remain unvalidated. The failed control leaves the unmatched estimates and their caveats in place.

Horizon dependence (F6b). Recomputed from the recorded first-unlock steps as cumulative success (a branch without an achievement before termination remains unsuccessful at later deadlines), at a short and a long horizon of the same rollouts. The wood HUD edit's effect in DreamerV3 grows from -0.215 at 8 steps to -0.707 at 25 ($p = 0.002$; Δ-IRIS -0.523 to -0.918, $n = 4$), and F3R2's recurrent damage from -0.014 at 4 steps to -0.281 at 13, in both cells. For stone nothing grows: DreamerV3's HUD effect shrinks (-0.096 to -0.041) and Δ-IRIS's is complete by 8 steps (-0.968). Every branch starts at the collection moment under a sustained intervention, so growth mixes opportunity (the tier takes several steps to reach), compounding divergence and any

forgetting; this is horizon dependence, not a memory-delay experiment. It replaces an earlier split at each agent's median gap since the last collection, which was degenerate because every moment is a first-collection frame.

K THE AUDIT: ERRORS FIXED, RERUNS AND VALIDATION

After the confirmatory runs, an item-by-item audit of the study found four implementation errors, unmatched controls and several over-readings. Its dated addendum (`runs/phase7/PROTOCOL_AUDIT_2026-09-19.md`) fixed the reruns and reanalyses before any was read. This appendix supersedes earlier drafts of the paper, which reported a null interaction between the code, recurrent and on-screen copies read as "the copies add", a HUD edit that "stops tool-making in every agent", a stone $z_t \times h_t$ substitution, and a single wood-specific dictionary feature.

Implementation errors. (1) The stone bases were rank-deficient (App. J); every basis is now rank-revealing, with its singular values and rank recorded. (2) Two diagnostic AUROC routines ranked tied scores by index order; one average-rank implementation replaced them (simulator outcomes are unaffected). (3) The HUD runs wrote condition names under two schemas (wood-specific and concept-generic); a loader now normalises them and fails on a missing condition. (4) The first fresh-learner run gave each arm its own initialisation, confounding initialisation with arm; it is superseded by a paired rerun (below).

Feature ranking. Every dictionary feature choice in this paper ranked features by the *raw* mean difference of a *proxy* activation, $\mathrm{ReLU}(x_{\text{std}}W_{\text{enc}})$, between label-positive and label-negative states; the trained dictionaries compute $\mathrm{TopK}_k\big(W_{\text{enc}}(x_{\text{std}} - b_{\text{dec}}) + b_{\text{enc}}\big)$ with $k = 32$. On the same design-split states, ranking instead by the standardized effect $d = (\mu^+ - \mu^-)/s_{\text{pooled}}$ on the true forward pass changes the state-selected pick at 16 of 34 sites, in 13 of 17 agents (usually to the original's second or third choice, once to its 31st). The original picks are stable under their own criterion (they rank first in a median 95% of 200 class-stratified bootstrap resamples), so the issue is the criterion rather than sampling noise. Under the true forward pass the original picks are active in only 18% of positive states (median; 3–58%), and no single feature separates the label well (AUROC 0.51–0.68; best feature 0.53–0.72). The pre-registered architecture experiment's candidate pool was built with the same proxy on the collection-event label: its top six overlap weakly with the robust ranking (Jaccard 0.00–0.71), and the recorded gate would remain a candidate at 6 of 34 sites. The gate was chosen by causal effect among those candidates, so the selection stands, but its search space was shaped by the proxy.

Restoration with the corrected pick (F2R). Re-running F2 with the corrected pick in the 13 agents where it changed keeps every sign but takes two of the four specificity contrasts below correction: rescue $+0.026$ (Holm $p = 0.064$; originally $+0.040$, $p = 0.029$) and right status above an unrelated patch $+0.020$ ($p = 0.087$; originally $+0.052$, $p = 0.030$); right above random $+0.035$ ($p = 0.007$) and right above wrong status $+0.426$ ($p = 8 \times 10^{-5}$) hold. The paper therefore makes no claim that a single dictionary feature is wood-specific. The right-versus-wrong effect does not depend on the dictionary—the patched subspace is defined by linear readouts—and is what supports the causal use of the availability subspace.

The primary repair experiment (F3R2). Two earlier repair runs, F3 and F3R, patched a six-dimensional object mixing availability with event and collection directions, and differed in donor matching and pairing; they gave rescues of $+0.211$ and $+0.116$ and are superseded. F3R2 uses the availability subspace of F2 and F7 and twelve conditions, each drawing at every step one wrong-status and one correct-status donor by the same rule from the same random stream (donor identities coincide until the rollouts diverge, since each rollout keys its donors on its own state): keep or damage each of z_t and h_t (the 2×2); after damaging both, repair h_t, z_t or both; positive controls that repair one site after damaging only that site; the repaired h_t routed to the actor only or to the transition only; and a single repair at the first step. Right-status donors matched every recorded flag in 99.98% of steps, wrong-status donors every flag of their flipped key in 99.97%. Two donor seeds give similar rescue estimates ($+0.311$ and $+0.316$; per-agent Spearman $+0.97$), showing stability within the new design; changes in basis and donor rules prevent attributing the earlier discrepancy to one factor. A null repair is read as a property of the site only if that repair's positive control succeeds: the h_t control does ($+0.298$, Holm $p = 0.001$), the z_t control does not ($+0.011$, $p = 0.10$). The

$z_t \times h_t$ interaction (-0.006, 90% CI $[-0.048, +0.036]$) is a different estimand from the effect of z_t damage once h_t is damaged (-0.076, $p = 0.002$), which is reported separately.

Validating the interventions actually deployed. A first validation refit the subspace on a capped subset and drew its own donors, so its numbers described a similar intervention rather than the deployed one; it is superseded. The F3R2 and F2 runs were rerun with logging only and reproduced the recorded rollout outcomes (17 of 17 agents in every set; at worst one differing rollout in about 2,000), so the logged patches are the ones applied. Along the actual trajectories, relative to $\|h_t\|$: the projection removes 0.042 (wood) and 0.036 (stone); a repair adds 0.056 (wood) and 0.046 (stone) to the projected state and moves h_t 0.042 and 0.039 from its pre-intervention value; F2's patches change the lesioned state by 0.041 (right status), 0.055 (wrong), 0.032 (unrelated) and 0.075 (random), and the gate lesion by 0.025; a code transplant changes 7.1 of the 8 slots. Offline, with the saved deployed basis, slots and donor rule on held-out states: projecting out W drops a fixed wood$\geq$1 readout from AUROC 0.999 to 0.783, but readouts retrained on projected states recover 0.978 (logistic and MLP); for stone $\geq$1, 0.948 fixed and 0.997 retrained; a random rank-matched projection changes nothing. A wrong-status h_t patch makes an independent readout report the donor's status (AUROC 0.85 wood, 0.79 stone; 77% and 57% of its decisions flip), a correct-status patch flips 1%. A code transplant moves an independent code readout little (wood $0.551 \rightarrow 0.526$): the code carries little decodable availability. Collateral of the h patch, measured the same way on every recorded label: a wrong-status patch flips the target readout's decisions in 77% (wood) and 57% (stone) of states, labels entailed by the target (wood$\geq$2, stone$\geq$2) in 17–23%, the faced-cell readout in 18.7% (wood) and 20.2% (stone), versus 14.9% and 15.6% for correct-status patches. Every other non-target label changes in at most 10% of states (median over agents; mean over non-target labels 5% and 8%). These differential changes do not establish content-independent collateral. W and F2's U have a largest principal cosine of 0.26 (median; 0.08–0.78).

Donor wall-status diagnostic. The deployed donor rule was evaluated offline on up to 3,000 held-out states per agent; it compares the binary `facing_wall` flag, not the full identity of the faced cell. In stone-holding receivers, correct-status agreement is 100% and wrong-status agreement 58.2% over the ten F4R agents. In wood-holding receivers, correct-status agreement is 78.5% over 17 agents: table placeability requires two wood, so one wood does not constrain wall status. All values are medians over agents. An earlier summary labelled $p^2 + (1-p)^2$, using only receiver wall frequency p, as chance agreement. That expression assumes independent draws with identical marginals; it is not a validated null for donors selected from conditional matching pools. We withdraw the chance interpretation, retain the reference values under an explicitly descriptive name in the diagnostic artifact, and make no chance-level claim. These offline rates do not measure donor agreement along the intervention trajectories or establish the cause of the behavioural gap.

The detector over a horizon. The DreamerV3 cascade in the pre-registered architecture experiment fits a logistic detector per tier achievement on $[z_t, h_t]$ with the threshold at the 95th percentile of negative design states, and scores whether any detector fires within the window. On real recorded 12-frame windows before the tier is unlocked, its per-state firing rate is 0.07–0.40 against a calibration rate of 0.05, and the cascade statistic is 0.566 (median over 17 agents) on windows with no tier event against 0.966 on windows with one. This diagnoses the detector on real windows; it does not decompose that experiment's imagined baseline into true and false firing, for which no ground truth exists.

The recorded cascades against what the lesion removes (F5). Using the exact recorded lesion vectors (dictionary gates) or a refit (the four probe gates, also excluded as a sensitivity check), and separating unrelated labels (stone, coal, sapling, stone pickaxe) from the wood tier's own descendants (wooden-tool possession), the joint-site cascade covaries with the one-step footprint on h_t ($\rho = +0.65$, $p = 0.004$) and with descendant loss ($+0.61$), not with event loss ($+0.06$) or availability loss ($+0.26$); leave-one-agent-out R^2 is at most 0.04. These are correlations across 17 agents, not a mediation analysis.

HUD edits by agent and achievement. The edit shows a zero count while the agent holds the resource (false information, not occlusion); the sham redraws the true count with the same pixel operation. Wood-tier achievement is mostly placing a table (sham: table 0.928, wooden pickaxe 0.705, wooden sword 0.014), and the edit lowers each (table 0.161, pickaxe 0.052). Four DreamerV3 agents keep 0.20–0.46 under the edit, so the on-screen count is load-bearing but its loss is not always decisive.

A fresh learner (F7). A new actor–critic is trained inside each frozen world model (imagination only, from the agent's own recorded start states; 16k steps at batch 128) with the recurrent wood subspace, a random subspace of equal rank, or nothing masked from what the actor and critics read. The budget was fixed on an intact-arm calibration that used agent 0's evaluation moments. Without that agent the estimates hold (wood − intact -0.046, wood − random -0.050), but wood − random is no longer significant after correction (Holm $p = 0.094$, against 0.047 with it), so we report both. Masking a random subspace of equal rank is equivalent to masking nothing within ± 0.05 (TOST $p = 0.002$); masking the wood subspace is not shown equivalent ($p = 0.49$). The first run gave each arm its own initialisation and is superseded; the paired rerun shares initial weights across arms (hash-verified), averages three learner seeds and applies the pre-set rule (intact fresh actor at least half the trained agent's rate) to the mean, keeping 8 of 10 rendered agents. In the hidden cell all seven fresh intact actors failed that rule at the frozen budget (0.05–0.30 against 0.66–0.87). A pilot at four times that budget on hidden agents 0 and 2 (intact arm, one learner seed; audit protocol B4) reached 0.060 and 0.399 against bars of 0.33 and 0.42. Agent 2 was still improving at its last evaluation (74 moments, two rollouts each), but by the pre-set rule no hidden-cell F7 was run, and the hidden-cell question is reported as unanswered.

Deviations from the protocols. Beyond the errors above: per-cell seed extensions were decided after the five-seed tests (§2); DreamerV3 dictionaries use $k = 32$, not the pre-registered 16 (App. F); DreamerV3's G1 detector is a calibrated probe rather than MI atoms (App. A); the Phase-7 z_t manipulation became a donor-code overwrite after F0 showed that projection does not change the sampled code; F2R, F3R2 and the paired F7 were added after the confirmatory runs, each with its own dated protocol; the F6b delay analysis was redefined; two descriptions in earlier drafts were wrong and are corrected here (the hidden cell does not occlude world wood; DreamerV3's gate selection used closed-loop, not teacher-forced, imagination); the first intervention validation described a refit rather than the deployed intervention and is replaced (above); one stone output from before the basis fix entered two appendix-level summaries because they were aggregated before the last rerun finished (re-aggregated; no number in this paper changed; the loader now rejects such outputs); and F4R was added with its own dated protocol after a second audit.

L PER-SEED RESULTS

Secondary, pre-lesion. Across the DreamerV3 cells the recurrent state carries $4.4\times$ more tool-tier signal in imagination than the stochastic latent (h_t 0.555 vs. z_t 0.125, $p = 5.9 \times 10^{-11}$), as P3 predicts, observed representationally before any lesion.

Table 5: G1 per cell. Δ-IRIS: top-2 MI codebook atoms (measured null 0.020); DreamerV3: calibrated probe (null 0.05 by construction). 100 burn-in moments per seed from 3000-episode probe corpora (1000 in hidden cells). Both arms $\approx 10^7$ environment steps.

cell	seeds	mean	pass
Δ-IRIS · rendered (prereg draw)	0.630, 0.563, 0.333, 0.496, 0.820	0.568	3/5
Δ-IRIS · rendered (re-draw, 10 seeds)	0.647, 0.357, 0.400, 0.543, 0.590, 0.470, 0.307, 0.790, 0.177, 0.480	0.476	4/10
Δ-IRIS · hidden (7 seeds)	0.097, 0.213, 0.393, 0.110, 0.278, 0.050, 0.207	0.193	0/7
DreamerV3 · rendered (12 seeds)	0.577, 0.567, 0.577, 0.433, 0.610, 0.567, 0.523, 0.657, 0.570, 0.530, 0.410, 0.570	0.549	10/12
DreamerV3 · hidden (8 seeds)	0.660, 0.457, 0.520, 0.553, 0.573, 0.533, 0.623, 0.507	0.553	7/8

Table 6: DreamerV3, every seed. Cascade = baseline − lesioned at each site for the causally selected gate (g), random (r) and the decodability-selected probe (f, joint site only). The last three columns describe the seed's own variance-matched null at the joint site: matched directions rolled out (n), the gate's rank-based empirical p, and its standardised effect z (undefined where all draws coincide).

cell	seed	G1	pass	n_{test}	probe rank	z_t g / r	h_t g / r	both g / r	both f	null n	p_{emp}	z
rendered	0	0.577	✓	128	1	+.003 / −.011	+.021 / −.005	+.036 / −.012	+.036	4	0.200	+3.4
rendered	1	0.567	✓	216	2	−.002 / +.005	+.271 / −.005	+.274 / +.001	+.025	38	0.026	+20.0
rendered	2	0.577	✓	98	3	+.005 / +.000	+.111 / +.001	+.122 / +.006	+.000	15	0.062	+23.0
rendered	3	0.433	×	188	2	+.014 / −.004	+.092 / −.023	+.096 / −.020	+.043	48	0.020	+5.3
rendered	4	0.610	✓	72	2	+.009 / +.006	+.015 / +.006	+.016 / +.013	+.019	18	0.158	+1.4
rendered	5	0.567	✓	400	2	−.006 / −.007	+.014 / +.000	+.017 / +.013	+.007	34	0.029	+3.7
rendered	6	0.523	✓	400	7	−.002 / +.001	+.191 / −.002	+.199 / −.002	−.014	34	0.029	+69.0
rendered	7	0.657	✓	57	1	+.004 / −.007	+.015 / +.006	+.023 / +.007	+.023	7	0.125	+1.5
rendered	8	0.570	✓	226	3	+.009 / −.002	+.044 / +.008	+.023 / +.016	+.008	28	0.069	+2.0
rendered	9	0.530	✓	80	1	+.011 / +.013	+.070 / +.011	+.070 / +.000	+.070	6	0.143	+12.0
rendered	10	0.410	×	113	2	+.001 / −.004	−.038 / −.014	−.010 / −.016	+.032	33	0.441	-0.0
rendered	11	0.570	✓	400	1	+.011 / −.003	+.023 / −.004	+.032 / −.001	+.032	6	0.143	+4.2
hidden	0	0.660	✓	334	6	+.015 / +.006	+.075 / +.009	+.083 / +.004	+.017	46	0.021	+7.7
hidden	1	0.457	×	400	7	−.000 / −.001	+.135 / +.007	+.159 / −.002	+.020	37	0.026	+7.1
hidden	2	0.520	✓	400	4	+.007 / +.004	+.156 / +.014	+.156 / +.005	+.039	17	0.056	+17.4
hidden	3	0.553	✓	327	6	−.002 / +.002	+.012 / −.007	+.015 / −.003	+.009	27	0.036	+3.1
hidden	4	0.573	✓	354	7	+.001 / +.001	+.328 / +.008	+.323 / +.004	+.007	40	0.024	+56.7
hidden	5	0.533	✓	387	6	+.004 / +.004	+.130 / +.000	+.130 / +.003	+.010	32	0.030	+26.9
hidden	6	0.623	✓	282	7	+.048 / +.010	+.119 / +.006	+.089 / +.008	+.007	34	0.029	+13.8
hidden	7	0.507	✓	337	3	+.001 / −.003	+.118 / −.007	+.117 / −.002	+.009	42	0.023	+25.2

Table 7: Δ-IRIS rendered, every seed, 400 test / 120 selection moments. CL = closed-loop, TF = teacher-forced. G1 given for the pre-registered draw and the re-draw; seeds 5–9 were trained after the pre-registration and have one draw.

seed	G1 prereg	G1 re-draw	baseline	gate lift-rank	CL gate	CL lift	CL CAV	CL random	TF gate	TF random
0	0.630	0.647	0.652	5	+.012	+.017	+.024	−.004	+.005	−.000
1	0.563	0.357	0.389	12	+.018	+.008	+.007	+.012	+.010	+.001
2	0.333	0.400	0.407	5	+.375	−.018	+.000	−.021	+.008	−.001
3	0.496	0.543	0.517	7	−.012	−.028	−.003	−.031	+.003	+.002
4	0.820	0.590	0.578	2	+.014	+.195	+.012	+.021	+.001	+.000
5	–	0.470	0.502	6	−.008	+.004	−.005	+.007	+.001	+.000
6	–	0.307	0.328	5	+.035	+.022	+.006	+.022	+.013	−.001
7	–	0.790	0.757	2	−.010	−.031	−.023	+.003	+.002	+.000
8	–	0.177	0.203	11	−.004	+.005	+.005	+.011	+.016	−.000
9	–	0.480	0.463	5	−.027	−.023	+.009	+.002	+.001	−.000

M EXPERIMENT REGISTRY

Table 8: Every Phase-7 experiment and its status, generated by `build_registry.py`. Only primary experiments support claims; superseded ones are kept as history and are never pooled with their replacements.

ID	Status	Object	Horizon, endpoint	Supports
F0	diagnostic	legacy gate (raw/std), label-selected SAE features, random	readout AUROC loss	the legacy gate carries the collection event, not availability
F1-real	primary	legacy gate (as recorded or refit probe)	25 real steps; wood-tier achievement (mostly place_table), simulator	behavioural impact
F1-imagined	primary (measurement susceptibility only)	legacy gate	12 imagined steps; calibrated detector (horizon FPR measured: see detector_fpr)	measurement susceptibility; no cross-endpoint subtraction
F2	primary for correct-vs-incorrect only; SAE specificity withdrawn (F2R)	W_avail (rank 4); control U not rank-matched (rank 3 in 12/17 agents); gate or label-selected SAE lesion	25 real steps; wood-tier achievement, simulator	the availability subspace is causally used
F2R	primary (robustness: withdraws specificity)	W_avail; SAE lesion re-picked by true-forward d	25; as F2	single-feature specificity not robust
F3	superseded by F3R2 (object differs; donors unpaired across conditions)	W6 (availability + event + collection), z donor slots	13 real steps; wood-tier achievement	—
F3R	superseded by F3R2 (object W6; coarse donor keys)	as F3	13; as F3	—
F3R2	primary (repair, route)	W_avail (identical to F2/F7), z slots	13 real steps; concept-tier achievement, simulator	usable restoration; route
F3R2-stone	exploratory	W_avail(stone), rank-revealing	13; stone-tier achievement (mostly place_stone)	concept dependence
F3-stone / F3R-stone	superseded (implementation error: rank-deficient basis)	stone basis with an arbitrary direction (duplicate label)	13; —	—
F4-wood	primary	HUD wood cell (false zero count); h: W_avail	25 real steps; wood-tier achievement, by achievement	the visible copy is load-bearing (large reduction, not abolition)
F4-stone	primary (HUD); h conditions rerun with the rank-3 basis	HUD stone cell; h: W_avail(stone)	25; stone-tier achievement (mostly place_stone)	architecture-stratified: Delta-IRIS depends on the stone count, DreamerV3 barely
F5	exploratory (descriptive only)	exact legacy vectors (SAE gates); refit (probe gates)	Spearman / LOO R^2	cascade covaries with footprint and descendant loss; not causal

Continued on next page

Table 8 continued

ID	Status	Object	Horizon, endpoint	Supports
F6b	exploratory (horizon dependence, not memory delay; replaces a degenerate median-gap split)	re-analysis of F4 and F3R2 outputs (recorded first-unlock steps)	8 vs 25 (F4); 4 vs 13 (F3R2); cumulative success by step	—
F7	superseded by F7-paired (initialisation confounded with arm)	W_avail masked from actor+critic input	25; wood-tier achievement	—
F7-paired	primary design; conclusion INCONCLUSIVE (8 of 10 rendered agents pass the competence rule)	as F7	25; as F7	adaptation to restricted inputs not established: wood-intact n.s. and not equivalent within 0.05; wood-random significant only with the calibration agent
F7 hidden pilot	exploratory (competence pilot, audit protocol B4); neither pilot passed	intact arm only	25; as F7	whether hidden-cell F7 can be attempted
F3R2-instr	diagnostic (reproduces F3R2; measures the applied patches)	as F3R2 (deployed objects saved)	13; as F3R2 + applied change per step	the actual F3R2 patch sizes
F2-instr	diagnostic (reproduces F2 gate; measures the applied patches)	as F2 (recorded U construction)	25; as F2 + applied change per step	the actual F2 patch sizes
F4R	primary for the stone content question (PROTOCOL_AUDIT2)	HUD stone cell x W_avail(stone): keep / drop / right-status / wrong-status donor	25 real steps; stone-tier achievement (mostly place_stone)	h's stone content is used when the HUD is falsified (replicated at 3 donor seeds); patch size and faced-cell matching unresolved
F4R-nm	FAILED magnitude-matching control for F4R (PROTOCOL_AUDIT4: matching criterion not met; contrasts reported under imperfect matching)	as F4R + donors selected to match the paired patch norm	25; as F4R	does not separate content from patch size
Validation (deployed)	diagnostic (offline, deployed objects)	deployed W_avail, z slots, DonorRule donors	retrained readouts; h/z patch content	projection removes directions, not information; what the patches carry
Donor faced-cell check	diagnostic (the donor key omits the faced cell)	deployed donor rule	faced-cell agreement, receiver vs donor	wall-status matching differs by donor status; exact correct matching constrains it at stone >= 1 or wood >= 2; no chance-level claim
SAE audit	diagnostic	SAE features	rank, bootstrap stability, true-forward firing	—

Continued on next page

Table 8 continued

ID	Status	Object	Horizon, endpoint	Supports
Intervention validation v1	superseded by the deployed-object validation and the instrumented reruns (not the deployed intervention)	W refit on a capped subset; own donors	retrained readouts, patch norms	—
Detector FPR	diagnostic	calibrated tier detector (legacy recipe)	12 real frames; horizon-level false positives	—

N PROVENANCE AND REPRODUCTION RECORDS

The initial predictions are recorded at tag `prereg-v1`; dated protocols under `runs/phase7/` govern the follow-ups. The provenance index contains 572 experiment-output entries covering 555 distinct files. A generating commit is recorded for 234 files; 142 of these were produced from working trees with unrecorded uncommitted changes. The other 321 files record no commit. Sixty-four files also carry a source manifest listing modified files and hashes of imported source files. A keep-alive operation reset output modification times on the scratch filesystem on 2026-09-22. The index therefore checks content hashes against its previously committed version; for 494 affected files, job attribution uses the latest log naming the file. Hashes and job identifiers cannot reconstruct missing source changes, checkpoints or corpora; missing commits are not inferred. Instrumented reruns reproduce the outcomes of the runs they re-execute. Tables and figures use committed aggregates; `paper/src/figures_v3.py` regenerates the six data figures from these and the published decode summary; `paper/src/fig_setup.py` draws Fig. 1 from committed recorded frames. App. K records corrections, superseded analyses and protocol deviations. Code, configurations and aggregate results accompany the paper.