BEYOND PRESENT RECALL: AUDITING FUTURE PLASTIC RESPONSE UNDER RECONSTRUCTIVE WRITEBACK

Anonymous authors
Paper under double-blind review

ABSTRACT

When neural memory architectures feed reconstructed content into plastic writeback, using stored memories alters the substrate for later queries. Within one controlled recurrent executor, matched interventions establish that prespecified present observables and writeback trajectories do not uniquely determine future response. Structural routing and runtime retrieval state causally reshape memory-wise response: the runtime contrast persists across identical stored traces, ruling out different written histories as the explanation in this executor. Common-state evaluation provides an expected attribution consistency check. Supporting stages show that reconstructive writeback selectively redistributes future memory quality, resolving excess memory-wise dispersion while aggregate excess remains unresolved at the primary comparison. At a fixed bank and schedule, a frozen predictor forecasts held-out quality changes across independent error directions, with post-freeze decomposition attributing this association to its direction-averaged component while stream-specific residuals do not transfer. In a separate differentiable system, post-hoc individual-memory calibration tests under oracle-residual inputs show known trace decay accounts for most calibration gain over an unweighted local baseline. Externally, visual memory replicates response redistribution and frozen prediction, whereas native same-checkpoint focal-layer test-time training preserves centered ordering and sign despite absolute common-mode failure. These findings motivate auditing memory-wise changes under specified writeback and retrieval conditions, alongside present recall and aggregate scores.

1 INTRODUCTION

Neural memory is increasingly implemented not as a read-only store, but as internal state that continues to change during inference. In fast-weight layers, recurrent cells, and test-time-training models, for example, internal parameters or hidden contexts update as new tokens or observations arrive (Sun et al., 2025; Behrouz et al., 2025). In architectures where retrieved or reconstructed content enters subsequent plastic updates, that writeback directly modifies the parameters or states accessed by later queries. A memory that successfully reconstructs an item at the present step may therefore alter the internal environment on which subsequent read operations depend.

Evaluating this dynamic requires tracking a complete episode: querying existing memories, reconstructing or retrieving stored content, applying a specified writeback update, and subsequently re-querying the original memory set (Figure 1). Measuring future scores against an experiment-specific matched reference isolates how that particular writeback trajectory changed their measured quality. Here present recall scores leave an immediate gap: how well a memory reconstructs content right now need not determine its future change under writeback. Furthermore, aggregate change scores present a second, distinct gap: a small or zero mean shift across all items can mask substantial heterogeneous redistribution, averaging out opposing improvements and degradations among individual memories.

Prior research analyzes related phenomena through update interference, gradient influence functions, and hidden learning-state dynamics in models sharing similar initial input–output behaviors (Koh & Liang, 2017; Pruthi et al., 2020; Lopez-Paz & Ranzato, 2017; Riemer et al., 2019; Natra-

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.

jan & Fitzgerald, 2025; Ger & Barak, 2026; Wang, 2026). Here the research question targets the specific process: under reconstructive writeback, what determines the signed changes of individual stored memories, and what structural or runtime factors drive this redistribution? Where the error-to-quality map is locally differentiable, we examine this sensitivity using standard first-order Jacobians relative to the matched reference, treating the operator as an analytical description rather than a new mathematical framework.

Within one controlled recurrent executor, matched structural-routing and retrieval-state interventions establish that prespecified present observables and writeback trajectories do not uniquely determine future response. Supporting stages first resolve selective memory-wise redistribution alongside unresolved aggregate excess at the primary comparison. At a fixed bank and schedule, a frozen predictor predicts memory-wise changes under held-out error directions, with post-freeze decomposition attributing this association to its direction-averaged component. In a separate differentiable system under oracle-residual inputs, post-hoc individual-memory calibration tests show that known trace decay accounts for most calibration gain over the unweighted local baseline. Externally, visual memory replicates response redistribution, whereas native same-checkpoint focal-layer test-time training preserves centered rank and sign despite absolute common-mode failure.

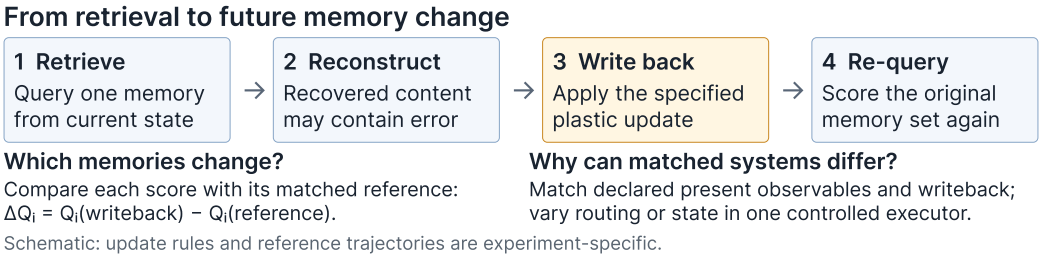


Figure 1: **Episodic writeback trajectory and audit questions.** Comparing re-queried outcomes against an experiment-specific matched reference isolates individual signed future changes ΔQ_i, distinguishing them from present recall and aggregate shifts. Within the controlled executor, prespecified present observables and writeback are matched while structural routing or runtime state differs. Architectures and reference trajectories remain experiment-specific.

2 Problem Setup: Future Plastic Response in Writable Memory

Reconstructive writable memory. We consider a writable-memory system with N individually queryable memories, where reconstructed content may enter plastic writeback with scale $\eta > 0$. Writeback inputs are defined by event errors $e_t \in \mathbb{R}^d$ stacked into a trajectory $e = \mathrm{vec}(e_1, \ldots, e_T) \in \mathbb{R}^{Td}$ relative to a matched, experiment-specific reference trajectory. Controlled analyses vary e directly, whereas endogenous settings evaluate retrieval-generated errors. Let $Q_i^{(h)}$ denote the prespecified, experiment-specific empirical recall/support score for memory i at horizon h, computed within an individual system, seed, or window rather than pooled across seeds (Appendix A). The memory-wise score vector and its relative change are

$$Q^{(h)}(e) = \big(Q_i^{(h)}(e)\big)_{i=1}^{N}, \qquad \Delta Q^{(h)}(e) = Q^{(h)}(e) - Q^{(h)}(0).$$

Present-matched denotes agreement on prespecified present observables within tolerance, not complete input–output equivalence.

Reconstructive-plastic response. When the error-to-quality map is differentiable at $e = 0$, score changes obey the local first-order expansion

$$V^{(h)} := \frac{1}{\eta} D_e Q^{(h)}(0), \qquad \Delta Q^{(h)}(e) = \eta V^{(h)} e + o(\|e\|).$$

Here $V^{(h)}$ is the standard local scaled Jacobian at the reference; its i-th row maps small writeback errors to memory i's future score change. The model-based operator $V_{\mathrm{prop}} := \frac{1}{\eta} D_e Q_{\mathrm{final}}(0)$ differentiates through known downstream plastic dynamics. The non-differentiable mismatch-gated architecture instead uses an ϵ-conditional susceptibility statistic (Section 4.1).

Population redistribution and hidden determinants. Suppressing h, centering via $P := I - \frac{1}{N}\mathbf{1}\mathbf{1}^\top$ isolates deviations from the across-memory common mode. The population redistribution statistic $R(e)$ measures the mean squared centered change, approximated via $G := V^\top PV$:

$$R(e) := \frac{1}{N}\|P\Delta Q(e)\|_2^2 \approx \frac{\eta^2}{N} e^\top Ge.$$

Thus V captures individual score changes, while G governs centered redistribution across memories. Redistribution reflects residual dispersion rather than conservation or transfer of memory quality. The excess-dispersion metrics $E_{\mathrm{rec,SD}}$ and E_{SD} are differences of across-memory standard deviations, distinct from R. In the controlled recurrent executor of Section 5, we write $V = V(A, u)$, where A is structural routing of memory and context through writable computation and u is runtime retrieval state governing expression or propagation of a fixed plastic trace.

3 Reconstructive Writeback Selectively Redistributes Future Memory Quality

Does imperfect reconstructive writeback produce differential future change across memories beyond that of a matched clean trajectory? We measure excess memory-wise dispersion alongside aggregate mean support and accuracy (Figure 2). These endpoints distinguish heterogeneous change from aggregate degradation.

3.1 Controlled reconstruction error induces selective future change

We use a separately registered, scale-decoupled design to isolate reconstruction-error effects from changes caused by writeback scale alone. Here $e_t = \epsilon d_t$ with $\|d_t\|_2 = 1$. At $\epsilon = 0$, state, weights, scale, support, and accuracy have zero cross-η difference over the tested scales (Figure 2A). Appendix B.1 documents the rejected predecessor and its clean-scale confound.

For each seed and η, the error condition is paired with its same-seed, same-η clean trajectory. We define $E_{\mathrm{rec,SD}}$ as the across-memory standard deviation of support change in the error condition minus that of its matched clean trajectory. At the preregistered primary slice $(\eta = 0.10, \epsilon = 1.0)$,

$$E_{\mathrm{rec,SD}} = 0.019978, \qquad 95\%\,\mathrm{CI} = [0.017148,\ 0.023508],$$

with all 16 seed-averaged effects positive (Figure 2B).

The effect is differential rather than an aggregate forgetting result. At the same slice, mean-support excess is 0.002707 (95% CI $[-0.000447, 0.005677]$) and future-accuracy excess is 0.008518 (95% CI $[-0.000977, 0.016927]$); neither interval excludes zero (Figure 2C). Thus aggregate excess is statistically unresolved at this slice despite resolved memory-wise redistribution.

3.2 Replication across implementations and endogenous error

The phenomenon survives changes in writable-memory implementation (Figure 2D). A native fast-weight/delta-write memory yields $E_{\mathrm{SD}} = 0.00071310$ (95% CI $[0.00062973, 0.00080958]$) across 512 aggregate cells and 12,288 memory-level observations; the count includes two stored error geometries at $\epsilon = 0$. A nonlinear recurrent mismatch-gated Hebbian memory yields $E_{\mathrm{rec,SD}} = 0.00912332$ (95% CI $[0.00598438, 0.01300789]$); all 16 seed estimates are positive. Full cell counts and recurrent audits are in Appendix B. Architecture-specific support scales preclude magnitude comparisons.

Finally, we remove externally imposed error directions. Imperfect retrieval generates the reconstruction used for ordinary self-writeback; a matched oracle arm instead writes back the corresponding target memory, with a frozen no-writeback reference. Self relative to oracle-writeback produces $E_{\mathrm{SD}} = 0.00594735$ (95% recovered-summary CI $[0.00434354, 0.00765131]$), positive in all 16 seeds. Selective redistribution therefore persists when the discrepancy is generated endogenously by the memory system itself. These experiments establish the phenomenon; we next ask whether a frozen response predicts its memory-wise pattern.

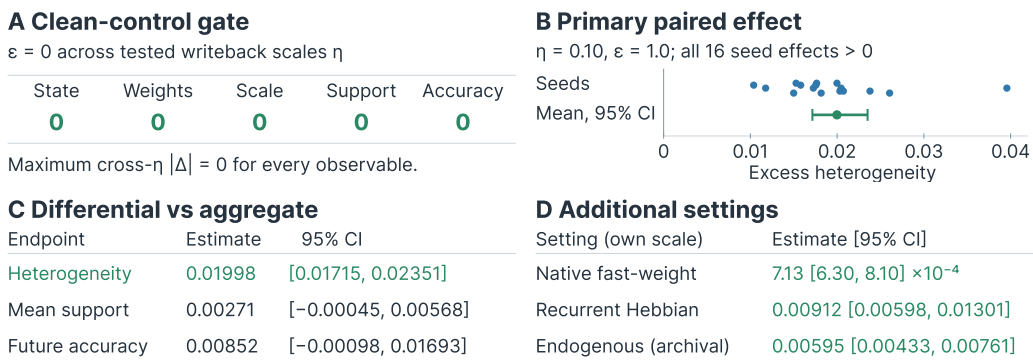


Figure 2: **Reconstructive writeback selectively redistributes future memory quality. (A)** The scale-decoupled error-free trajectory is invariant across tested writeback scales. **(B)** Reconstruction error produces positive excess memory-wise heterogeneity at the preregistered primary slice. **(C)** The differential endpoint resolves while both aggregate endpoints remain statistically unresolved. **(D)** The effect replicates in two further writable-memory implementations and with endogenous reconstruction error. The endogenous interval in (D) is archival; the main text labels the recovered-summary interval (Appendix I).

4 Frozen Response Structure Predicts Future Plastic Change

We examine two prediction questions: whether susceptibility repeats under independently resampled error directions, and whether propagation through known dynamics calibrates finite-horizon response. Each test fixes the response before evaluation; its conditional inputs are specified below.

4.1 Direction-averaged susceptibility predicts under resampled errors

We evaluate whether a susceptibility computed with one isotropic error-direction stream predicts future quality changes under an independent stream for the same memories. We analyze the 32-dimensional mismatch-norm-gated recurrent architecture (B-OOS; Appendix C.1), distinct from the 12-dimensional causal executor of Section 5. Here, updates are stored in a writable fast-weight matrix H, and eight bipolar unit memories $\mathcal{M} = \{p_i\}_{i=1}^{8}$ are replayed across three balanced cue cycles (24 events total), with the memory bank and cue schedule held fixed within each seed.

Let $B_i = \nabla_H Q_i(H; h_T^0)|_{H=0}$ denote the local score gradient with respect to H at the clean reference holding the terminal clean state h_T^0 fixed; B_i measures local score sensitivity rather than a propagated total derivative. Let Γ_t^A be the reference-trajectory write increment at event t with update scale η removed, constructed at error magnitude ϵ using the independent audit direction stream D_A. Summing Frobenius matrix inner products over all 24 events defines the scalar susceptibility and its prediction:

$$S_i^{\text{local}}(D_A) = \sum_{t=1}^{24} \langle B_i, \Gamma_t^A \rangle_F = S_i^{(0)} + S_i^{\text{dir}}(D_A),$$
$$\widehat{\Delta Q_i^B} = \eta S_i^{\text{local}}(D_A). \tag{1}$$

The first relation in Equation 1 defines an exact ϵ-conditional statistic, separated from the derivative operator V of Section 2 because the norm gate is non-differentiable at $e = 0$; the second defines a prediction rather than an equivalence to observed outcomes. The direction-averaged term $S_i^{(0)}$ is independent of the direction stream but still depends on the bank, schedule, clean dynamics, and ϵ.

Crucially, the full predictor $\widehat{\Delta Q_i^B}$—including the bank, schedule, D_A, and ϵ—was frozen before evaluating any D_B outcome; no D_B errors, updates, or final traces entered its construction. The exact decomposition into $S_i^{(0)}$ and directional residual was performed only post-freeze; $S_i^{(0)}$ was not independently prefrozen. This protocol tests cross-direction prediction at a fixed bank and schedule, not cross-bank, cross-schedule, or cross-system generalization (Appendix C.1; Appendix I).

Across the fixed 1,152 held-out memory-level rows (inference over 16 seed clusters, $n = 16$), the frozen response gives Pearson $= 0.97388$ (seed-cluster 95% CI $[0.96311, 0.99101]$), Spearman $= 0.99732$ ($[0.99543, 0.99832]$), and uncentered sign accuracy $= 99.392\%$ ($[98.177\%, 100\%]$; Figure 3A). Exact post-freeze decomposition gives the no-direction baseline Pearson $= 0.973874$, Spearman $= 0.997320$, and the same sign accuracy; its seed-cluster intervals are reported in Appendix C.1.

All nine fixed $\eta \times \epsilon$ cells pass individually (within-cell Pearson $[0.956, 0.998]$, Spearman $[0.992, 0.998]$), ruling out a between-cell pooling explanation. The majority-negative baseline is 90.017%, so uncentered sign accuracy improves by 9.375 points; correlations remain primary. A post-freeze robustness reanalysis that demeans prediction and outcome within each seed–η–ϵ cell retains Pearson $= 0.9671$, Spearman $= 0.9949$, and sign accuracy $= 98.44\%$; a within-cell memory-permutation null gives $p < 10^{-4}$. Its through-origin slope $= 0.5432$ and no-fit $R^2 = 0.2737$ make the remaining finite-horizon gain error explicit.

In contrast, after subtracting the common prediction from both sides, $\eta S_i^{\mathrm{dir}}(D_A)$ versus $\Delta Q_i^B - \eta S_i^{(0)}$ gives Pearson $= -0.137880$, Spearman $= -0.090918$, and sign accuracy $= 48.438\%$; all three seed-cluster intervals include their null values. The directional term accounts for at most 9.51×10^{-6} of full-predictor variance in any fixed $\eta \times \epsilon$ cell. The post-freeze decomposition therefore attributes the cross-stream association to direction-averaged susceptibility. The stream-specific residual shows no predictive evidence across the independent streams.

On the same 1,152 rows, pooled (Pearson, Spearman) scalar baselines are error norm $(-0.243, -0.241)$, update norm $(-0.267, -0.253)$, current support $(0.249, 0.583)$, current margin $(0.238, 0.541)$, and gradient norm $(-0.654, -0.768)$. Gradient norm is strongest, yet remains unsigned; its within-cell $|r| = 0.901$–0.940 and $|\rho| = 0.858$–0.862 remain below the signed susceptibility ($r = 0.956$–0.998, $\rho = 0.992$–0.998), which also predicts sign without fitting (Appendix C).

4.2 Finite-Horizon Calibration in a Differentiable System

We examine conditional finite-horizon calibration in a separate differentiable ungated recurrent system (B-THEORY-002; Appendix C.2). Evaluated around the frozen reference trajectory, full propagation $V_{\mathrm{prop}} = \frac{1}{\eta} D_e Q^{\mathrm{final}}(0)$ follows perturbations through subsequent decay and state-dependent writes. The local baseline also differentiates the terminal six-step probe while holding reference presynaptic states fixed. Both predictions condition on specified error inputs, reserving the realized self-writeback outcome for evaluation.

Across the 384 individual-memory observations, the same-base local response remains miscalibrated (through-origin slope 0.491198, no-fit $R^2 = -0.570522$), whereas full propagation restores calibration (slope 1.007022, no-fit $R^2 = 0.997154$; Figure 3B–C). A new, separately recorded post-hoc control weighting local event contributions c_{it} (including η) by known decay, $\sum_{t=0}^{23} 0.93^{23-t} c_{it}$ without fitted parameters, yields slope 1.012217 and no-fit $R^2 = 0.996966$, capturing most calibration gain. Full propagation yields a slightly higher pooled no-fit R^2, but the paired seed-cluster 95% interval for its RMSE gain over decay weighting includes zero (Appendix C.3). This decay attribution applies strictly to this 384-row individual-memory comparison, as decay baselines were not evaluated on the subsequent population redistribution or finite out-of-sample direction tests.

Using $G = V_{\mathrm{prop}}^{\top} P V_{\mathrm{prop}}$, we select high-response, smallest-positive, numerical-null, and independent isotropic directions from the frozen geometry before observing their consequences; error norms are matched at 0.05, 0.10, or 0.20. Across 1,440 non-null direction–amplitude rows, population prediction has Pearson $= 0.999970$, slope $= 0.99710$, and no-fit $R^2 = 0.999931$. Thus equal-magnitude errors can redistribute memory differently according to their alignment with the frozen response geometry.

Outside this grid, the oracle-base residual is recorded along a prespecified oracle/reference trajectory after V_{prop} is frozen and before the self arm produces $Q_{\mathrm{self}} - Q_{\mathrm{oracle}}$. Error-trajectory norms average 4.064 (range 2.931–4.676). Across 48 seed–writeback cells, propagated population prediction remains calibrated (Pearson $= 0.99823$, slope $= 1.01428$, no-fit $R^2 = 0.99616$), whereas the same-base local predictor does not (slope $= 0.23355$, no-fit $R^2 = -14.68$). This prediction conditions on the known oracle trajectory and its residuals; it does not forecast those residuals from present memory. Input independence from the self outcome does not assert that numerical predictions were

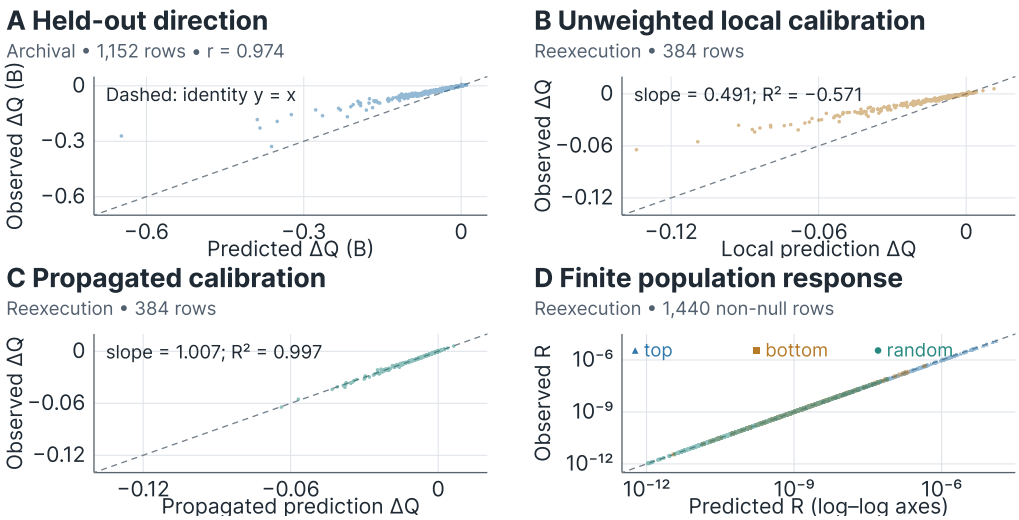


Figure 3: **Direct tests of memory-wise and population response prediction. (A)** Frozen susceptibility predictions versus observed ΔQ_i^B across 1,152 preserved frozen held-out memory rows (16 seed clusters) in the mismatch-norm-gated system. **(B–C)** Comparison of unweighted local (**B**) and propagated (**C**) predictions against observed ΔQ on identical axes across 384 rows (16 seeds) from the existing separately labelled reexecution of the separate differentiable ungated system (B-THEORY-002). On these 384 rows, a separately recorded post-hoc decay-weighted control (not shown) also achieves near-unit calibration (slope 1.012, no-fit $R^2 = 0.99697$; Section 4.2, Appendix C.3). **(D)** Predicted versus observed population redistribution R on log-log axes across 1,440 non-null rows from that ungated reexecution (144 numerical nulls excluded), stratified by top, smallest-positive (bottom), and random directions. Dashed lines indicate identity; reported slopes are diagnostic through-origin fits, and no-fit R^2 evaluates fixed, unadjusted predictions. Plotted points reflect nested observations rather than independent replicates (Appendix I).

historically saved first (Appendix C.2). The tested agreement establishes finite-trajectory calibration in this setting.

We next isolate how routing and retrieval state alter response in one controlled recurrent executor.

5 Causal Response Changes in a Controlled Recurrent Executor

Which features can change response while present observations and writeback are held fixed? We isolate structural routing A and retrieval-time state u in one controlled recurrent executor. Each intervention preserves the declared present/writeback ledger needed for causal isolation. Matching covers prespecified observables, not arbitrary input–output equivalence; the structural intervention changes off-memory-manifold routing. Construction, matching, mutable-state, and washout ledgers are in Appendices D–E. For $X \in \{V, G\}$, $d_X(X_1, X_2) := \|X_1 - X_2\|_F / [\frac{1}{2}(\|X_1\|_F + \|X_2\|_F)]$; $d_V(A)$ and $d_V(u)$ denote the structural and runtime pairs. The matched-ledger contrasts estimate changes in measured V under the declared A or u intervention.

5.1 Structural routing reshapes future susceptibility

To isolate how routing and retrieval state govern response, we examine a 12-dimensional recurrent executor storing six target memories. The space is partitioned into a memory subspace spanned by orthonormal basis $U \in \mathbb{R}^{12\times 6}$ with projector $P_H = UU^\top$, and an orthogonal complement spanned by the orthonormal basis $N \in \mathbb{R}^{12\times 6}$ ($U^\top N = 0$) with projector $P_N = NN^\top$.

The writable state is a plastic matrix A_T^{pl}, updated from matched exogenous write trajectories independently of routing or runtime conditions. Memory i is retrieved over 12 steps via

$$h^{(\ell+1)} = \tanh\big[(W_{\mathrm{struct}} + K_u + 0.80A_T^{\mathrm{pl}})h^{(\ell)} + b_i\big],$$

initializing from $h^{(0)} = 0$ under fixed cue drive b_i.

Structural routing A alters cross-subspace transmission via orthogonal matrices $O_{A/B}$, whereas runtime retrieval state u_j introduces an unwritten modulation R_j acting on the same block:

$$W_{A/B} = \beta P_H + \lambda P_N + \gamma U O_{A/B} N^\top, \qquad K_{u_j} = c\,U(O_A R_j)N^\top \quad (j = 1, 2).$$

Fixed scalar parameters follow Appendix D. Structural comparisons set $K_u = 0$ and switch W_A/W_B; runtime comparisons fix W_A and vary K_u. Both interventions thus act through the complement-to-memory block $U(\cdot)N^\top$ rather than disjoint supports. In the nonlinear recurrent experiment, the two structural conditions match present content and context decoding, clean reconstruction and trajectories, reconstruction error, plasticity strength and complete plastic increments, together with relevant recurrent-scale diagnostics. Despite this matching,

$$d_V(A) = 0.12085, \qquad \frac{\|V_B\|_F}{\|V_A\|_F} = 0.999958,$$

where d_V denotes normalized Frobenius operator difference. The operator angle is 6.93°, and the induced population geometry also changes, $d_G(A) = 0.17188$. Thus the intervention primarily rotates future susceptibility rather than producing a high-gain versus low-gain system (Appendix D). Figure 4A shows the paired memory-wise response differences.

Legal permutation and signed-permutation symmetries reproduce equivalent dynamics, but do not relate the measured A/B pair. An orthogonal conjugacy of its linear components is not a symmetry of the full elementwise-$\tanh$ architecture: its mean normalized nonlinear equivariance error is 0.182388 over 5,000 frozen off-manifold probes (Appendix D.1). These controls exclude the tested coordinate explanations for the pair.

5.2 Retrieval-time state changes the expression of an identical trace

At fixed A, u is a runtime retrieval state acting on off-memory-manifold propagation. The two control conditions match prespecified tested present observables, reconstruction, recurrent scale, reconstruction error at every write, plasticity coefficient, every plastic increment, and the final stored trace.

In particular, every plastic increment ΔA_t and the final stored plastic trace are identical across conditions:

$$\max_t \|\Delta A_t(u_1) - \Delta A_t(u_2)\|_\infty = 0.$$

Because the plastic matrix is the only mutable cross-write state and u remains an unwritten runtime condition, this exact update identity rules out different written updates or persistent stored traces as explanations for the between-arm response difference in this executor. Nevertheless,

$$d_V(u) = 0.0291115, \qquad \frac{\|V_2\|_F}{\|V_1\|_F} = 1.00001844,$$

again indicating a primarily orientational response change (Appendix E); the paired memory-wise response differences are shown in Figure 4B.

To separate stored-trace differences from state-conditioned expression, we probe both identical traces under common u_0 with fresh transient state $h = 0$. Neither plastic trace is erased. The audit verifies zero between-arm difference (Figure 4C; Table 2):

$$\max_i |\Delta Q_i^{u_1 \to u_0} - \Delta Q_i^{u_2 \to u_0}| = 0, \qquad |R^{u_1 \to u_0} - R^{u_2 \to u_0}| = 0.$$

With identical traces and deterministic common-state evaluation, this vanishing contrast serves as an expected attribution check. Paired runtime interventions across identical written histories attribute response differences to state-conditioned expression rather than stored-trace divergence, while equal within-arm responses remain nonzero. Together with structural routing, these audits establish

$$\boxed{V = V(A, u)}.$$

Within this controlled executor, prespecified present observables and writeback trajectories do not uniquely determine future response.

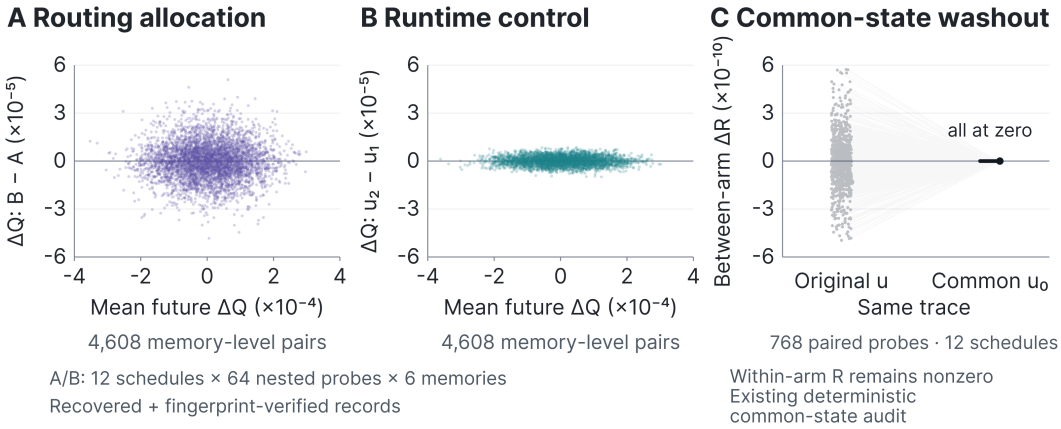


Figure 4: **Controlled interventions on future response. (A–B)** Difference versus mean memory-wise response on identical axes across 4,608 matched rows per panel (12 schedules $\times$ 64 nested probes $\times$ 6 memories, not independent replicates; recovered, fingerprint-verified records). **(A)** Structural routing alters signed response. **(B)** Runtime state alters expression of an identical stored trace. **(C)** Existing deterministic audit across 768 paired probes compares signed population contrast $R_{u_2} - R_{u_1}$ before versus after common-u_0 evaluation (paired lines and horizontal jitter for visibility). All post-washout contrasts vanish exactly, with stored traces unchanged and within-arm R nonzero. Evidence is confined to one controlled recurrent executor with matched prespecified present observables and writeback (Appendix I).

5.3 Weak interaction permits a regime-specific additive description

The complete 2×2 map $V_{ij} = V(A_i, u_j)$ did not meet the preregistered strong-interaction criterion: $\|I_V\|_F / \operatorname{mean}_{ij} \|V_{ij}\|_F = 6.76052 \times 10^{-5}$ (0.00676%), where $I_V = V_{22} - V_{21} - V_{12} + V_{11}$ and the denominator averages the four corner norms. Across schedules and complementary-factor values, structural, runtime, and interaction distances lie in $[0.1172, 0.1256]$, $[0.02909, 0.02918]$, and $[6.51 \times 10^{-5}, 7.02 \times 10^{-5}]$. The registered extension was not run.

After the full map was frozen, these results motivated the regime-specific description $V(A, u) \approx V_0 + V_A(A) + V_u(u)$. The subsequent three-corner reconstruction has residual exactly $-I_V$; its downstream evaluation is therefore a consistency check, not independent factorization evidence (Appendix F).

6 Replication Beyond the Controlled Causal System

We next test memory-wise redistribution and response prediction in visual writable memory and a pretrained TTT executor. These experiments examine selected response findings; the matched A/u interventions remain confined to Section 5.

6.1 Out-of-family writable visual memory

CIFAR-10 content, frozen ImageNet-pretrained ResNet18–PCA64 features, and soft writable slots replace the synthetic setting; imperfect retrieval generates error endogenously. At this single operating point, 12 frozen seeds—the independent units, not pooled memory rows—yield $E_{\mathrm{SD}} = 0.00208136$ (95% CI $[0.00182977, 0.00236436]$; $12/12$ positive), while future accuracy remains 1.0. A frozen response predicts held-out exogenous individual and population change with no-fit $R^2 = 0.998071$ and 0.998089. Conditioning on the realized self-path residuals also predicts endogenous change ($R^2 = 0.998371$); the residuals themselves are not forecast from initial memory. The experiment thus resolves support changes at an operating point with perfect future accuracy.

6.2 Relative response on held-out TTT windows

We evaluate a focal reconstruction score (not language-model token likelihood) across 12 deterministic, non-overlapping WikiText windows excluded from B-MODERN-001, using the same pretrained TTT-MLP-125M checkpoint and layer-6 executor (Sun et al., 2025). Windows serve as statistical units. P2 (exogenous errors) and P3 (native-target discrepancy) share an initial writable state and 48-token future segment, conditioning predictions on extracted future inputs and specified target discrepancies; P2 tests perturbations with amplitude $a = 0.05\|y^{\mathrm{nat}}\|_2$ (Appendix G.2). Under P1 (native writeback relative to matched oracle), all $12/12$ windows show positive selective rewriting ($E_{\mathrm{SD}} = 0.0204227$, 95% CI $[0.0181281, 0.0233622]$); held-out P2 yields individual and population-redistribution no-fit $R^2 = 0.805572$ and 0.952933.

For the native-target discrepancy, P3b separately centers predictions and outcomes within each window. It gives Pearson 0.740237, Spearman 0.742245, sign 0.747396, no-fit $R^2 = 0.0599703$, and slope 0.514485. Uncentered absolute/common-mode P3a fails (sign 0.345052, no-fit $R^2 = -4.42951$); no correction is fitted. A window-cluster bootstrap preserves rank/sign but gives centered no-fit R^2 CI $[-0.200, 0.274]$ (Appendix G). The retained ordering and sign concern deviations from each window's respective mean; calibrated centered magnitude is not established.

7 Related Work

Writable states, retrieval-dependent modification, interference, derivative influence, and forgetting forecasts are well established (Miconi et al., 2018; Schlag et al., 2021; Sun et al., 2025; Behrouz et al., 2025; Osan et al., 2011; Lopez-Paz & Ranzato, 2017; Riemer et al., 2019; Koh & Liang, 2017; Pruthi et al., 2020; Jin & Ren, 2024; Hacohen & Tuytelaars, 2025). Closely related work demonstrates that loss-invisible structure alters subsequent adaptation and models future-response variations on present-behavior fibers (Ger & Barak, 2026; Wang, 2026), while representation flux and consolidation risks further motivate post-writeback auditing (Kazanskii, 2026; Zhang et al., 2026). Distinct from these precursors, our focus is the signed future change of already stored memories under reconstructive writeback, with causal routing and retrieval-state dissociations isolated within a controlled executor.

8 Discussion and Limitations

Within one controlled executor, matched prespecified present observables and writeback trajectories do not uniquely determine future response. Runtime interventions produce response differences across identical stored traces, ruling out different written histories as the between-condition explanation. These between-condition response differences can reflect state-conditioned expression rather than altered persistent content. These results motivate reporting signed item-level changes alongside aggregate scores, pairing mean change with centered redistribution under declared retrieval states. This audit serves as a support diagnostic; predicting or preventing task failure remains unestablished.

Complete A/u causal evidence is confined to one controlled executor with prespecified matching, without multi-amplitude or cross-architecture mechanism tests, and its additive description is regime-specific. B-OOS supports direction-averaged susceptibility under error-direction resampling at a fixed bank and schedule. Visual response findings reflect one operating point with perfect future accuracy. In the same-checkpoint focal-layer TTT model, native tests preserve centered rank and sign, but absolute common-mode prediction fails and calibrated centered magnitude is unestablished. Differentiable analyses remain local, first-order, and conditional on known dynamics and error inputs, with the norm-gated system analyzed separately. Estimation with incomplete dynamics and response-aware design remain open.

AI USE STATEMENT

Generative AI tools were used as research-assistance tools throughout this project. They assisted with developing and refining conceptual and mathematical framing, proposing and refining hypotheses, designing and reviewing experimental protocols, implementing and reviewing experimental code, interpreting intermediate and final results, conducting literature searches and prior-art comparisons, drafting and editing portions of the manuscript, and creating or revising scientific figures.

All reported numerical results were obtained from executed experimental code rather than generated by a language model. AI-assisted code, analyses, mathematical arguments, and scientific claims were checked against the corresponding experimental outputs, frozen protocols, numerical validation tests, and primary literature before inclusion. Negative results and preregistered stopping or decision rules were retained rather than modified to match AI-generated suggestions. Scientific figures created or revised with AI-assisted tooling were manually checked against the reported results.

Generative AI tools were not treated as authors and did not replace author responsibility for experimental decisions, scientific claims, or the final manuscript. We take responsibility for all content of this work, including text, code, analyses, and figures produced with AI assistance.

REPRODUCIBILITY STATEMENT

Experimental definitions, frozen evaluation protocols, causal controls, held-out prediction procedures, numerical derivative audits, negative results, experimental chronology, and external-model validation are documented in the appendix. The submitted anonymous supplementary ZIP includes `CLAIM_TO_ARTIFACT_INDEX.md`, which routes each claim to frozen evidence, and `PROVENANCE_AND_SCOPE.md`, which distinguishes archival originals, recovered/fingerprinted records, and reference reconstructions. Missing bytes are disclosed; no reconstruction is represented as a byte-identical original.

REFERENCES

Ali Behrouz, Peilin Zhong, and Vahab Mirrokni. Titans: Learning to memorize at test time. In *Advances in Neural Information Processing Systems*, 2025.

Yoav Ger and Omri Barak. Learning reveals invisible structure in low-rank rnns. *arXiv preprint arXiv:2605.04115*, 2026.

Guy Hacohen and Tinne Tuytelaars. Predicting the susceptibility of examples to catastrophic forgetting. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 21546–21569. PMLR, 2025.

Xisen Jin and Xiang Ren. What will my model forget? forecasting forgotten examples in language model refinement. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 22145–22159. PMLR, 2024.

Maksim A. Kazanskii. Geometry of forgetting: Representation flux in continual learning. *arXiv preprint arXiv:2608.15854*, 2026.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 1885–1894. PMLR, 2017.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Thomas Miconi, Kenneth O. Stanley, and Jeff Clune. Differentiable plasticity: Training plastic neural networks with backpropagation. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 3559–3568. PMLR, 2018.

Maanasa Natrajan and James E. Fitzgerald. Stability through plasticity: Finding robust memories through representational drift. *Proceedings of the National Academy of Sciences*, 122(45): e2500077122, 2025.

Remus Osan, Adriano B. L. Tort, and Olavo B. Amaral. A mismatch-based model for memory reconsolidation and extinction in attractor networks. *PLOS ONE*, 6(8):e23113, 2011.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, and Gerald Tesauro. Learning to learn without forgetting by maximizing transfer and minimizing interference. In *International Conference on Learning Representations*, 2019.

Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. Linear transformers are secretly fast weight programmers. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 9355–9366. PMLR, 2021.

Yu Sun, Xinhao Li, Karan Dalal, Jiarui Xu, Arjun Vikram, Genghan Zhang, Yann Dubois, Xinlei Chen, Xiaolong Wang, Sanmi Koyejo, Tatsunori Hashimoto, and Carlos Guestrin. Learning to (learn at test time): Rnns with expressive hidden states. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 57503–57522. PMLR, 2025.

Qinyou Wang. Fiber fingerprints of hidden learning-state dynamics. *arXiv preprint arXiv:2608.15976*, 2026.

Dylan Zhang, Yanshan Lin, Zhengkun Wu, Yihang Sun, Bingxuan Li, Dianqi Li, and Hao Peng. Useful memories become faulty when continuously updated by llms. *arXiv preprint arXiv:2605.12978*, 2026.

Appendix roadmap. Appendix A defines experiment-specific scores and reporting units; Appendix B documents redistribution controls; Appendix C gives B-OOS information flow, baselines, and propagation; Appendices D–F audit the declared-observable matching, mutable state, washout, and additive-analysis chronology; Appendix G reports scoped external validation; and Appendices H–I record negative boundaries and artifact provenance.

A Experimental Discipline and Reporting Conventions

All confirmatory experiments were separated from implementation development and calibration. When a frozen experimental contract was used, calibration was restricted to disjoint seeds or prespecified present-only diagnostics. Confirmatory outcomes were never used to change primary endpoints, redefine predictors, modify evaluation directions, or rescue failed gates.

Throughout the paper, *present-matched* denotes agreement with the declared present observable profile under the specified tolerance. It does not imply complete behavioral equivalence over arbitrary inputs. All reported no-fit R^2 values compare unadjusted predictions with observations, without post-hoc slope, intercept, or rescaling.

Operational quality scores. Q_i is always computed for memory i within one experimental system, seed, or text window; "support" is shorthand for that experiment-specific empirical recall score, not a population pooled across seeds. In B-ORTHO-001/B-DECOUPLED-001 it is the cue average of $\mathrm{clip}[(\cos(z, p_i) + 1)/2, 0, 1]$. In B-ARCH-002 it is the temperature-5 softmax probability assigned to stored value v_i from the cosine similarities of normalized Wk_i to all stored values. In B-ARCH-003, B-ENDOGENOUS-001, B-THEORY-002, and B-OOS-DIRECTION-001 it is the temperature-6 correct-memory softmax probability from cosine similarities between the full-cue recurrent probe and all stored bipolar patterns. In the shared A/u mechanism system it is the normalized target projection $Q_i = H_i^\top h_i/(H_i^\top H_i)$ after a 12-step memory-specific probe. In B-EXTERNAL-001 it is the correct-slot mass $[\mathrm{softmax}_{\mathrm{slots}}(4Mp_i)]_i$. In B-MODERN-001R it is $Q_i(\theta) = -\frac{1}{2}\,\mathrm{mean}_{h,d}(\widehat{r}_i(\theta, XK_i) - (XV_i - XK_i))^2$. Here XK_i, XV_i are the frozen focal-layer projected key and value representations of probe token i; $\widehat{r}_i$ is the layer-normalized TTT-MLP reconstruction of their residual. The writable state θ contains the two MLP weight matrices and biases per head, and the mean is over 12 heads and 64 coordinates per head. This is a read-only focal reconstruction score, not language-model token likelihood; the update and error-coordinate mapping are in Appendix G.2.

Excess-dispersion labels. $E_{\mathrm{rec,SD}}$ and E_{SD} retain the frozen experiments' labels for excess across-memory dispersion. Each is a difference of across-memory standard deviations, not the standard deviation of a paired difference. Controlled error conditions use their matched zero-error arm; endogenous tests compare self/native and oracle changes relative to the same initial or no-writeback reference. These experiment-specific references are retained rather than replaced by a single pooled endpoint. Neither label is identical to R, which is the mean squared centered paired quality change.

Population redistribution is evaluated after removing the across-memory common mode:

$$P = I - \frac{1}{N}\mathbf{1}\mathbf{1}^{\top}.$$

Bootstrap confidence intervals resample the declared experimental units: systems/seeds, or windows for TTT. Pooled memory/token rows are not treated as independent replicates. Response magnitudes are compared across architectures only when explicitly stated.

B Selective Redistribution: Causal Controls and Replications

B.1 Clean-trajectory control

The initial orthogonal intervention (B-ORTHO-001) revealed a confound: at zero reconstruction error, changing writeback scale alone caused a weight-scale collapse from approximately 0.2344 to 0.0032. Therefore the observed drift could not be attributed specifically to reconstruction-error-mediated plastic change. This experiment was retained as a frozen NO-GO result.

The subsequent decoupled control (B-DECOUPLED-001) was designed so that the clean trajectory remained invariant across writeback scales. Across $1040/1040$ factorial cells and $16/16$ seeds, the maximum difference across η at $\epsilon = 0$ in state, weights, scale, support, and accuracy was zero.

At the preregistered primary slice,

$$E_{\mathrm{rec,SD}} = 0.019978, \qquad 95\%\,\mathrm{CI} = [0.017148, 0.023508],$$

with all 16 seed-level effects positive. Mean-support and future-accuracy intervals crossed zero, supporting selective redistribution rather than generic aggregate degradation.

B.2 Native fast-weight replication

The native fast-weight replication (B-ARCH-002) replaces the controlled construction with a key-local delta-write associative memory:

$$h_t = W_t k_j, \qquad \tilde{h}_t = h_t + \epsilon e_{j,t},$$
$$W_{t+1} = W_t + \eta(\tilde{h}_t - h_t)k_j^{\top}.$$

Unlike the initial construction, no corrupted-minus-clean proposal subtraction is used. Therefore the error-free condition produces an exactly zero update.

The frozen design uses 16 seeds, 24 memories, four replay cycles, and

$$\eta \in \{0, .01, .025, .05\}, \qquad \epsilon \in \{0, .05, .10, .20\}.$$

It records both error geometries even at $\epsilon = 0$, giving $16 \times 4 \times 4 \times 2 = 512$ aggregate cells and $512 \times 24 = 12{,}288$ memory rows; 256 would incorrectly collapse the stored geometry dimension at zero error.

The clean gate passes exactly for fast weights, norm, readout activity, support, margin, and accuracy. The primary heterogeneity endpoint is

$$E_{\mathrm{SD}} = 0.00071310, \qquad 95\%\,\mathrm{CI} = [0.00062973, 0.00080958].$$

All nine nonzero $\eta \times \epsilon$ cell means are positive. The absolute magnitude is not compared directly with other architectures because support scales differ.

B.3 NONLINEAR RECURRENT AND ENDOGENOUS-ERROR REPLICATION

The nonlinear recurrent replication (B-ARCH-003) evaluates whether the phenomenon persists beyond key-local associative memory. The model uses bipolar memories with mismatch-gated Hebbian plasticity:

$$H_{t+1} = 0.93H_t + \eta \frac{\|z_t - r_t\|_2 z_t q_t^\top}{32}.$$

Here and throughout, $\|\cdot\|_2$ denotes the Euclidean norm; the subscript is not an exponent.

The error-free trajectory satisfies $H_t = 0$ for all η, and the clean gate passes at tolerance 10^{-12}. The frozen design gives

$$E_{\mathrm{rec,SD}} = 0.009123, \qquad 95\%\,\mathrm{CI} = [0.005984, 0.013008],$$

with $16/16$ positive seed estimates.

The endogenous-error replication (B-ENDOGENOUS-001) instead uses ungated Hebbian writeback, as defined in Appendix C.2; it is not the norm-gated update above. Partial cues generate the model reconstruction $\hat{r}_t$, and self-writeback is compared with oracle writeback under matched cues, ordering, retrieval count, and memories. Using 16 seeds, eight memories, three replay cycles, and

$$\eta \in \{.01, .03, .05\},$$

self-writeback again produces larger across-memory heterogeneity:

$$E_{\mathrm{SD}} \approx 0.00595,$$

with bootstrap confidence excluding zero.

C RESPONSE ESTIMATION, PROPAGATION, AND NUMERICAL VALIDATION

C.1 HELD-OUT DIRECTION-RESAMPLING SUSCEPTIBILITY PROTOCOL

Throughout the paper, "B-OOS" abbreviates the supplementary artifact ID `B-OOS-DIRECTION-001`.

For each of 16 seeds, the audit uses the memory bank $\mathcal{M} = \{p_i\}_{i=1}^8$ of bipolar unit patterns in $D = 32$ and three independently permuted passes over them (24 events). Let $\Pi = [p_1^\top; \ldots; p_8^\top] \in \mathbb{R}^{8\times 32}$ be the pattern matrix, distinct from the population-centering projector P. With $W = 1.8\Pi^\top\Pi$, zero diagonal, $\alpha = .8$, cue gain $g_c = 3$, and five micro-steps, event dynamics are

$$h_t^{(\ell+1)} = \tanh\Big[(W + \alpha H)h_t^{(\ell)} + g_c p_{j_t} \mathbf{1}\{\ell < 2\}\Big].$$

Writing $r_t = h_t^{(5)}$, $\tilde{h}_t = h_t^{(4)}$, and letting $\psi_i(H; 0.1h_T)$ denote the full-cue probe initialized at $0.1h_T$, quality is

$$Q_i(H; h_T) = \left[\mathrm{softmax}\left(6\left[\frac{p_k^\top \psi_i(H; 0.1h_T)}{\|p_k\|_2 \|\psi_i(H; 0.1h_T)\|_2}\right]_{k=1}^{8}\right)\right]_i.$$

The packaged probe and gradient routines use 10^{-15} denominator stabilizers. Here r_t denotes an event state, whereas ψ_i denotes a memory-specific read-only probe. On the clean $H = 0$ trajectory, define $B_i = \nabla_H Q_i(H; h_T^0)|_{H=0}$. The memory bank and balanced cue schedule are fixed by the protocol; no bank or schedule resampling is claimed. The audit directions are 24 isotropic unit vectors generated from stream 2101; independently, the held-out directions use stream 2201. The exact predictor is equation 1, where $\|z_t^A - r_t^0\|_2 = \epsilon$. Hence its update and susceptibility split exactly as

$$\Gamma_t^A = \frac{\epsilon}{32} r_t^0 (\tilde{h}_t^0)^\top + \frac{\epsilon^2}{32} d_t^A (\tilde{h}_t^0)^\top = \Gamma_t^{(0)} + \Gamma_t^{\mathrm{dir}}(D_A),$$

$$S_i^{(0)} = \sum_t \langle B_i, \Gamma_t^{(0)} \rangle_F, \qquad S_i^{\mathrm{local}}(D_A) = S_i^{(0)} + S_i^{\mathrm{dir}}(D_A).$$

The held-out branch replays the nonlinear system under the same balanced schedule with

$$z_t^B = r_t^B + \epsilon d_t^B, \quad \Gamma_t^B = \frac{\|z_t^B - r_t^B\|_2}{32} z_t^B (\tilde{h}_t^B)^\top, \quad H_t^B = .93 H_{t-1}^B + \eta \Gamma_t^B,$$

and evaluates

$$\Delta Q_i^B = Q_i(H_T^B; h_T^B) - Q_i(0; h_T^0).$$

The original frozen information split is therefore

$$\underbrace{(\mathcal{M}, W, J, D_A, \epsilon) \to S_i^{\text{local}}(D_A) \xrightarrow{\times \eta} \widehat{\Delta Q_i^B}}_{\text{frozen predictor}} \quad \Big\| \quad \underbrace{(\mathcal{M}, W, J, D_B, \eta, \epsilon) \to \Delta Q_i^B}_{\text{held-out outcome}},$$

where J is the shared schedule and the branches meet only when metrics are computed. Neither D_B, Γ_t^B, H_T^B, nor ΔQ_i^B is available during predictor construction. The post-freeze analytic baseline is stricter:

$$\underbrace{(\mathcal{M}, W, J, \epsilon) \to S_i^{(0)} \xrightarrow{\times \eta} \widehat{\Delta Q_i^{B,(0)}}}_{\text{no-direction baseline}} \quad \Big\| \quad \underbrace{(\mathcal{M}, W, J, D_B, \eta, \epsilon) \to \Delta Q_i^B}_{\text{held-out outcome}}.$$

It uses neither stream 2101 nor stream 2201.

Scalar-baseline definitions. Let $\mathcal{T}_i = \{t : j_t = i\}$ denote the three audit events assigned to memory i, and let s_{ik}^0 be its clean probe's cosine similarity to pattern p_k. On each seed–memory–η–ϵ row, the frozen baseline columns are

$$\text{error norm}_i = \sum_{t \in \mathcal{T}_i} \epsilon = 3\epsilon, \qquad \text{update norm}_i = \sum_{t \in \mathcal{T}_i} \|\Gamma_t^A\|_F,$$

$$\text{current support}_i = Q_i(0; h_T^0), \qquad \text{current margin}_i = s_{ii}^0 - \max_{k \neq i} s_{ik}^0,$$

$$\text{gradient norm}_i = \|B_i\|_F.$$

Thus update norm sums unit-writeback proposal norms over that memory's own audit events: it is not $\|H_T\|_F$, not the norm of their sum, and is not multiplied by η. Gradient norm differentiates with respect to the plastic matrix H, not the input or full model parameters. All five use the clean/audit branch only and keep their original signs; none is fitted to the held-out outcomes.

The grid is $\eta \in \{.01, .03, .05\}$ and $\epsilon \in \{.05, .10, .20\}$, giving 128 seed–memory rows per fixed cell and 1,152 memory-level rows total. PASS requires pooled Pearson $> .4$, Spearman $> .4$, sign accuracy $> .65$, and at least 7/9 fixed cells with both Pearson and Spearman $> .3$. All nine cells pass. Error norm is constant within each fixed cell, so its within-cell correlation is undefined. Across the nine cells, Pearson ranges from 0.956068 to 0.997807, Spearman from 0.991926 to 0.997803, and sign accuracy from 0.992188 to 1.0. The pooled majority-sign baseline is 0.900174, versus frozen-predictor sign accuracy 0.993924. The five frozen scalar baselines were evaluated on the same 1,152 memory-level rows with their original signs and no fitting. Their pooled (Pearson, Spearman) values were error norm $(-0.242942, -0.240926)$, update norm $(-0.266860, -0.252685)$, current support $(0.249044, 0.582962)$, current margin $(0.238078, 0.541232)$, and gradient norm $(-0.654383, -0.768015)$, versus $(0.973880, 0.997323)$ for the frozen signed susceptibility. Within cells, error norm is undefined because it is constant; the remaining (Pearson, Spearman) ranges are update norm $([-0.188199, -0.174464], [-0.094139, -0.082671])$, current support $([0.314319, 0.401449], [0.611438, 0.658682])$, current margin $([0.305920, 0.376200], [0.563049, 0.610568])$, and gradient norm $([-0.940217, -0.901238], [-0.861890, -0.857781])$.

Across 1,152 held-out memory-level rows, with inference clustered over 16 seeds ($n = 16$),

$$\text{Pearson} = 0.974, \qquad \text{Spearman} = 0.997, \qquad \text{uncentered sign accuracy} = 99.4\%.$$

Resampling all rows of each seed as one cluster for 10,000 deterministic bootstrap replicates gives respective 95% percentile intervals $[0.963112, 0.991013]$, $[0.995434, 0.998318]$, and $[0.981771, 1.000000]$. Leave-one-seed-out ranges are $[0.971673, 0.987952]$, $[0.997144, 0.997726]$, and $[0.993519, 1.000000]$.

As an explicitly post-freeze robustness analysis, we demean prediction and outcome within each seed–η–ϵ group. The centered rows give Pearson 0.967077, Spearman 0.994879, sign accuracy

0.984375, through-origin slope 0.543163, no-fit $R^2 = 0.273657$, and RMSE 0.0176774. In 10,000 deterministic null replicates (RNG seed 51018), the centered predictor is permuted across the eight memories separately within every group. Null Pearson has mean 0.000101, 95% interval $[-0.180458, 0.213803]$, and maximum 0.412498; the one-sided empirical $p = (1 + \#\{r_{\text{null}} \geq r_{\text{obs}}\})/10001 = 1/10001 < 10^{-4}$. This robustness check was not part of the original frozen PASS criterion.

The post-freeze direction-component analysis reconstructs every stored $S_i^{\text{local}}(D_A)$ to maximum absolute error 5.33×10^{-15}; an independent NumPy reverse-mode implementation also recovers every stored gradient norm to 3.11×10^{-15}. The no-direction predictor $\eta S_i^{(0)}$ gives Pearson 0.973874, Spearman 0.997320, and sign accuracy 0.993924. Ten thousand seed-cluster bootstrap replicates (fixed RNG seed 51018) give respective 95% intervals $[0.962947, 0.991176]$, $[0.995472, 0.998316]$, and $[0.981771, 1]$.

For the residual audit we compare $\eta S_i^{\text{dir}}(D_A)$ with $\Delta Q_i^B - \eta S_i^{(0)}$ on the same frozen rows. Pearson is -0.137880 (95% CI $[-0.480544, 0.340087]$), Spearman is -0.090918 ($[-0.283600, 0.103031]$), and sign accuracy is 0.484375 ($[0.382812, 0.578125]$). Within every fixed $\eta \times \epsilon$ cell, the direction-specific term contributes at most 9.51×10^{-6} of full-predictor variance. This post-freeze analysis narrows the interpretation of the preregistered PASS: it supports direction-averaged susceptibility under the fixed bank/schedule and resampling, not cross-stream replication of a direction-specific residual; the latter is the residual-direction null.

C.2 Local and propagated response operators

Ungated recurrent system (B-THEORY-002). Section 4.2 uses the ungated endogenous memory of B-ENDOGENOUS-001, not the mismatch-norm-gated B-ARCH-003/B-OOS system. It has eight unit bipolar patterns in $D = 32$, three balanced passes ($T = 24$ writes), 16 seeds, and $\eta \in \{.01, .03, .05\}$. For $\Pi = [p_1^\top; \ldots; p_8^\top]$, set $W = 1.8\Pi^\top\Pi$ with zero diagonal. Each retrieval starts at $h_t^{(0)} = 0$ and uses a fixed mask m_t exposing 16 coordinates:

$$h_t^{(\ell+1)} = \tanh\Big[(W + .8H_t)h_t^{(\ell)} + 3(m_t \odot p_{j_t})\mathbf{1}\{\ell < 2\}\Big], \quad \ell = 0, \ldots, 5,$$

$$q_t(H_t) = h_t^{(5)}, \qquad \hat{r}_t(H_t) = \frac{h_t^{(6)}}{\|h_t^{(6)}\|_2 + 10^{-12}}.$$

The exogenous coordinate map writes $z_t = p_{j_t} + e_t$ and updates

$$H_{t+1} = .93H_t + \frac{\eta}{32}(p_{j_t} + e_t)q_t(H_t)^\top, \qquad H_0 = 0.$$

The oracle uses $e_t = 0$; the self arm instead writes its own $\hat{r}_t$. After all writes, a fresh full-cue probe for memory i gives normalized response $\psi_i(H_T)$ by the same six-step retrieval with $m = \mathbf{1}$, and $Q_i^{\text{final}} = [\text{softmax}(6\Pi\psi_i(H_T))]_i$. No factor $\|e_t\|_2$ gates this update. Differentiability is used locally around the audited reference trajectory, not asserted for every possible state of a normalized system. The propagated unit-writeback response is

$$V_{\text{prop}} = \frac{1}{\eta} D_e Q_{\text{final}}(0).$$

What the local comparison omits. Let H_T^{oracle} and q_t^{oracle} be the saved oracle terminal trace and presynaptic states. The packaged `local_same` comparison is exactly

$$B_i^{\text{oracle}} = \nabla_H Q_i(H)|_{H=H_T^{\text{oracle}}}, \qquad (V_{\text{local}})_{i,t} = \frac{1}{32}(B_i^{\text{oracle}} q_t^{\text{oracle}})^\top,$$

$$\widehat{\Delta Q_i}^{\text{local}} = \frac{\eta}{32} \sum_t e_{\text{base},t}^\top B_i^{\text{oracle}} q_t^{\text{oracle}}.$$

Here $(V_{\text{local}})_{i,t}$ denotes the row block for event t. This comparison differentiates through the final six-step probe, but freezes the oracle presynaptic states and does not propagate each write perturbation through the later write sequence, including its decay and state-dependent retrieval. There is no explicit per-event downstream decay weight in this local sum. In contrast, V_{prop} differentiates the complete recurrence above. The difference is therefore not "no recurrent differentiation" versus "recurrent differentiation".

Post-hoc decay-weighted local control. To isolate known trace decay from subsequent state-dependent writes, we evaluate a separately recorded post-hoc control (formulated 2026-09-22; not historically preregistered). For zero-based write events $t \in \{0, \ldots, 23\}$, let

$$c_{it} = \frac{\eta}{32} e_{\mathrm{base},t}^{\top} B_i^{\mathrm{oracle}} q_t^{\mathrm{oracle}},$$
$$\widehat{\Delta Q}_i^{\mathrm{local}} = \sum_{t=0}^{23} c_{it}, \qquad \widehat{\Delta Q}_i^{\mathrm{decay}} = \sum_{t=0}^{23} 0.93^{23-t} c_{it}.$$

Here c_{it} includes η; the final event has weight $0.93^0 = 1$, with all decay factors fixed by the known update rule rather than fitted post hoc. All three predictors evaluate the identical 384 observations (16 seeds $\times$ 3 learning rates $\times$ 8 memories) using the same oracle-base errors and realized $Q_{\mathrm{self}} - Q_{\mathrm{oracle}}$ targets. Inputs were recreated via packaged scientific functions; the original responses and predictions were checked against the previously reexecuted 384-row table rather than raw historical artifacts (Appendix I).

Conditional inputs and execution order. The oracle-base error is $e_{\mathrm{base},t} = \hat{r}_t(H_t^{\mathrm{oracle}}) - p_{j_t}$. $Q_{\mathrm{oracle}} = Q_{\mathrm{final}}(0)$ and Q_{self} are the same terminal quality score evaluated in the two respective arms. The self residuals evolve on the self trajectory; we do not assert $Q_{\mathrm{self}} = Q_{\mathrm{final}}(e_{\mathrm{base}})$.

This derivative propagates through the complete known recurrent/plastic computation and is frozen first. The oracle/reference arm then generates the residual vector e_{base} and its terminal Q_{oracle}, followed by the self arm and its terminal Q_{self}. The runner forms the numerical prediction afterward, but neither V_{prop} nor e_{base} depends on the self outcome or its residual trajectory:

$$\widehat{\Delta Q} = \eta V_{\mathrm{prop}} e_{\mathrm{base}}, \qquad \Delta Q_{\mathrm{obs}} = Q_{\mathrm{self}} - Q_{\mathrm{oracle}}.$$

The recovered historical summary records validation of autograd derivatives against central finite differences with step 10^{-5} across all 48 seed–writeback cells. The maximum absolute discrepancy was

$$3.77079 \times 10^{-11},$$

and the maximum relative directional discrepancy was

$$4.35580 \times 10^{-7}.$$

The supplementary provenance note preserves the differing historical report values and the separately labeled reexecution results.

No clipping, smoothing, saturation correction, or parameter search was introduced.

C.3 Finite-horizon calibration

Across the 384 individual-memory observations (16 seeds $\times$ 3 learning rates $\times$ 8 memories), evaluated on identical oracle-base errors against the observed $Q_{\mathrm{self}} - Q_{\mathrm{oracle}}$ target, the unweighted local baseline preserves memory ordering but remains uncalibrated. The post-hoc decay-weighted local control captures most of the calibration gain, while full propagation adds modest additional refinement:

Method	Spearman ρ	Slope	$R^2_{\text{no-fit}}$
Unweighted local	0.992448	0.491198	-0.570522
Decay-weighted local	0.999497	1.012217	0.996966
Full propagated	0.999493	1.007022	0.997154

Reported slopes serve as through-origin diagnostics; $R^2_{\text{no-fit}}$ evaluates direct identity calibration without fitted rescaling. Defining gain as decay error minus full propagated error (where positive values favor full propagation), RMSE, computed as the square root of pooled MSE across all 384 rows, gives an RMSE gain of 1.65457×10^{-5} with a 95% percentile CI of $[-1.69003 \times 10^{-6}, 3.76462 \times 10^{-5}]$. The corresponding pooled MAE gain is 9.17494×10^{-6} with a 95% percentile CI of $[6.80591 \times 10^{-7}, 1.98872 \times 10^{-5}]$. At the seed level, full propagation yields lower MSE in 9 of 16 seeds and higher MSE in 7.

Confidence intervals are estimated via paired seed-cluster percentile bootstrap (20,000 draws, resampling 16 seed IDs with replacement while retaining all 24 nested rows per sampled seed, pairing methods and recomputing pooled errors on each draw, RNG seed `20260922`) to account for dependence within seeds. This post-hoc evaluation uses recovered original scientific functions; baseline outputs were checked against previously reexecuted original-source records (Appendix I). This comparison applies strictly to these 384 individual-memory observations; the subsequent population redistribution and finite out-of-sample evaluations do not evaluate a decay baseline.

At the population level, frozen equal-norm directions with

$$\|e\| \in \{0.05, 0.10, 0.20\}$$

give

$$\text{Pearson} = 0.999970, \qquad R^2_{\text{no-fit}} = 0.999931.$$

For each of the 16×3 seed–η cells, the frozen geometry supplies one top, one smallest-positive, one numerical-null, and eight independent isotropic directions. Each is evaluated at all three norms: $16 \times 3 \times 3 \times (1 + 1 + 1 + 8) = 1{,}584$ direction–amplitude rows. The pooled non-null statistic excludes 144 null rows and uses 1,440 rows; the independent seed units remain 16. The recovered runner and the separately labeled `experiments/B-THEORY-002/reexecuted/finite_oos.csv` verify this design and count. Historical row-level CSVs remain unavailable; the quoted historical point metrics are retained, not relabeled as newly recovered raw results.

For oracle-base reconstruction errors observed before the self outcome, propagated trajectory-conditional prediction remains calibrated:

$$\text{Pearson} = 0.99823, \quad \text{Spearman} = 0.99631,$$
$$\text{slope} = 1.01428, \qquad R^2_{\text{no-fit}} = 0.99616.$$

The corresponding same-base local predictor has

$$\text{slope} = 0.23355, \qquad R^2_{\text{no-fit}} = -14.68.$$

C.4 Non-differentiable recurrent boundary

The response operator is not assumed to exist for every writable-memory architecture. In B-ARCH-003, the mismatch gate contains $\|e\|_2$ at $e = 0$. Therefore a global vector Fréchet derivative and the induced quadratic form

$$G = V^\top P V$$

are not defined at that base point.

This architecture is therefore used as phenomenon and directional-prediction evidence rather than as evidence for a universal linear operator.

D Structural-Allocation Intervention Audit

Frozen contracts, rows, and analyses are under `experiments/B-ALLOC-002/` in the supplementary artifact.

The control generator is the fixed, non-random matrix

$$S = I_3 \otimes \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \in \mathbb{R}^{6\times 6}, \qquad S^\top = -S.$$

Thus $R_1 = \exp(.60S)$ and $R_2 = \exp(-.60S)$ are the frozen paired rotations used in both structural conditions; O_A remains fixed inside K_{u_j}. No future response was used to initialize S.

The structural intervention modifies only an off-memory-manifold routing component while preserving stored memories, labels, readouts, plasticity rules, and clean reconstruction trajectories. No future outcome was used to select structural conditions.

Present-only matching gives content and metadata accuracy 1.0, mean current support

$$0.999999999979,$$

Table 1: Construction and intervention ledger for the controlled A/u executor. Dimensions and constants are fixed before future-outcome evaluation.

Object	Dimension	Fixed status and role
Targets H; bases U, N	12×6 each	Six stored targets and their memory/null bases; identical in every arm
$P_H = UU^\top, P_N = NN^\top$	12×12	Fixed projectors; $\beta = .25, \lambda = .08, \gamma = .35$ fixed
O_A, O_B	6×6	Frozen structural routing choices; only this choice changes in the A intervention
R_1, R_2	6×6	Frozen controls $R_1 = \exp(.60S)$, $R_2 = \exp(-.60S)$; $c = .10$ in $K_{u_j} = cU(O_A R_j)N^\top$
$W_{A/B}, K_{u_1}, K_{u_2}$	12×12	Structural and runtime routing terms; stored targets, readouts, drive, and plastic rule fixed
Schedule/error bank	12 schedules; 64 errors each	Paired across arms and chosen without future outcomes; each error has 12 write events
Plastic trace A_t^{pl}	12×12	Reset to zero for each schedule/error arm; the only mutable object carried across writes

reconstruction RMSE approximately

$$1.49 \times 10^{-11},$$

and identical recurrent Frobenius norm, operator norm, and spectral radius. The paired error, writeback scale, and every plastic increment are also identical; hence $\max_t \|\Delta A_t^{\mathrm{pl}}(A_1) - \Delta A_t^{\mathrm{pl}}(A_2)\|_\infty = 0$. These checks concern the declared observable/state ledger in Table 1; they do not assert equivalence on arbitrary off-manifold inputs.

Across 12 independent balanced schedules,

$$\frac{\|V_B\|_F}{\|V_A\|_F} = 0.999958, \qquad d_V(A) = 0.12085,$$

with operator angle 6.93°. At the induced population level,

$$\frac{\|G_B\|_F}{\|G_A\|_F} = 0.999994, \qquad d_G(A) = 0.17188,$$

with angle 9.86°.

The normalized structural distance ranges from 0.117239 to 0.125593 across schedule and runtime-state margins, rather than being driven by a single schedule.

Thus the intervention changes response orientation while preserving global response magnitude. This is a measured-executor proof-of-principle rotation; no operational threshold or cross-architecture magnitude law is established.

D.1 Symmetry and basis-equivalence controls

Permutation and signed-permutation transformations applied consistently to states, drives, errors, traces, and readouts reproduce identical future dynamics to numerical precision. Therefore the audit detects legal coordinate symmetries.

The structural pair is not related by such a transformation. Although the linear components admit an exact orthogonal conjugacy, this mapping is not a symmetry of the complete nonlinear elementwise-$\tanh$ architecture. On clean retrieval preactivations, the nonlinear-gate diagnostic is

$$L_\phi = \frac{\mathbb{E}\|\tanh(x) - x\|_2}{\mathbb{E}\|x\|_2} = 0.188091,$$

with saturation fraction 0. Separately, 5,000 frozen off-manifold probes give

$$\mathbb{E}\left[\frac{\|\tanh(Tx) - T\tanh(x)\|_2}{\|\tanh(Tx)\|_2 + 10^{-15}}\right] = 0.182388$$

for the tested orthogonal conjugacy T. These tests exclude the evaluated coordinate explanations within this architecture; they do not exclude arbitrary nonlinear diffeomorphisms.

E RETRIEVAL-TIME CONTROL AND WASHOUT

The complete mutable-state and washout ledger is under `experiments/B-CONTROL-001/`.

Table 2: Mutable-state and timing ledger for the retrieval-state intervention.

State/object	Persistence or reset	Paired-arm check
Runtime condition u	Fixed during trace generation; never written	Only K_{u_1} versus K_{u_2} changes
Transient probe state h	Initialized at zero for every retrieval probe	Not carried between memory probes or across arms
Plastic trace A_t^{pl}	Reset to zero per schedule/error arm; then carried across 12 writes	Every increment and final trace identical; maximum difference 0
Schedule, errors, scale	Fixed and paired	Same order, error tensor, $\eta = .03$, decay, and plastic gain
Washout evaluation	After trace equality is checked, both traces are probed under common u_0 with fresh $h = 0$	Maximum memory-wise and population-response differences both 0

At fixed structural allocation, the retrieval-time intervention (B-CONTROL-001) changes control state while matching present memory function, reconstruction, reconstruction error, writeback scale, and all exogenous plastic increments. Therefore the complete final plastic trace is identical. In the audited executor, this plastic matrix is the only mutable state carried across writes; retrieval control is a fixed runtime condition rather than a written state.

Global recurrent scale is also matched:

$$\|W_2\|_F / \|W_1\|_F = 1,$$

with equal operator norm and spectral radius.

Despite identical traces,

$$d_V(u) \approx 0.02911, \qquad d_G(u) \approx 0.04112,$$

and the frozen response predicts the resulting control-conditioned future differences. Across schedule and structural margins, normalized runtime-state distance is confined to $[0.029094, 0.029175]$.

The washout experiment tests whether this difference represents stored memory change or state-conditioned expression. After generating the identical exogenous plastic trace, both arms are evaluated under the same neutral control u_0. A dedicated deterministic common-u_0 audit checks this zero result from the identical trace and common evaluator; its raw output, summary, and hashes are packaged. The audit reconstructs each arm's trace from frozen inputs using the production trace builder, rather than loading saved per-pair matrices. Only the transient probe is restarted: neither plastic trace is erased or relearned. The 768 saved schedule–error rows have equal, nonzero within-arm redistribution (for example, the first row has $R^{u_1 \to u_0} = R^{u_2 \to u_0} = 2.40553 \times 10^{-9}$). The zeros below are between-arm contrasts, not zero within-arm response:

$$\max_i |\Delta Q_i^{u_1 \to u_0} - \Delta Q_i^{u_2 \to u_0}| = 0,$$

and

$$|R^{u_1 \to u_0} - R^{u_2 \to u_0}| = 0.$$

Within this executor's state ledger, retrieval-time control changes expression or propagation of an identical trace. The washout supports only the bounded claim that no persistent control-dependent difference remains after equalization; it establishes no operational threshold or cross-architecture magnitude law.

F FACTORIAL INTERACTION AND ADDITIVE CONSISTENCY

The prospective interaction record and post-freeze additive audit are under `experiments/B-INTERACT-001/` and `experiments/B-THEORY-003/`, respectively.

F.1 Interaction hypothesis and chronology

The factorial analysis was performed before evaluating the simplified additive description.

The preregistered interaction experiment (B-INTERACT-001) evaluated the complete

$$2 \times 2$$

response map:

$$V_{ij} = V(A_i, u_j).$$

The hypothesis that allocation would substantially modulate the individual-memory effect of retrieval-time control did not pass its gate. The measured interaction was

$$I_V = V_{22} - V_{21} - V_{12} + V_{11},$$

with

$$\frac{\|I_V\|_F}{\mathrm{mean}_{ij} \|V_{ij}\|_F} = 6.76052 \times 10^{-5}.$$

Its schedule-wise range is $[6.50939 \times 10^{-5}, 7.02131 \times 10^{-5}]$.

The preregistered endogenous interaction extension was not performed.

Only after the complete factorial result was frozen was the simpler near-additive description evaluated.

F.2 Three-corner consistency check

The closure analysis reconstructs the fourth response corner as

$$\widehat{V}_{22} = V_{21} + V_{12} - V_{11}.$$

This is a zero-refit consistency test, not blind discovery of the factorization rule. The additive description was selected after observing the complete factorial map.

By construction,

$$\widehat{V}_{22} - V_{22} = -I_V,$$

so the fourth-corner residual is algebraically identical to the factorial interaction and is not counted as independent evidence.

Applying $\widehat{V}_{22}$ to frozen held-out error directions gives

$$R^2_{\mathrm{no-fit}} = 0.999466, \qquad \text{uncentered sign accuracy} = 99.13\%$$

for individual-memory future change, and

$$R^2_{\mathrm{no-fit}} = 0.999805$$

for population redistribution.

The two reported residual percentages use different declared denominators. For schedule s, let $I_V^{(s)} = V_{22}^{(s)} - V_{21}^{(s)} - V_{12}^{(s)} + V_{11}^{(s)}$. Then

$$\frac{1}{12} \sum_{s=1}^{12} \frac{\|I_V^{(s)}\|_F}{\frac{1}{4} \sum_{i,j} \|V_{ij}^{(s)}\|_F} = 0.00676052\%,$$

whereas the residual relative to the raw A_2 control effect is

$$\frac{1}{12} \sum_{s=1}^{12} \frac{\|I_V^{(s)}\|_F}{\|V_{22}^{(s)} - V_{21}^{(s)}\|_F} = 0.232071\%.$$

Thus the numerator is identical but the second denominator is the much smaller within-A_2 control main effect; the factor-of-34 difference is a scale change, not a conflicting residual.

F.3 Population-level interaction induced by response geometry

Because

$$G = V^\top PV,$$

an approximately additive response family induces the cross-term

$$V_A^\top PV_u + V_u^\top PV_A.$$

Without fitting a separate G-level interaction model, this term explains

$$99.9584\%$$

of observed squared-Frobenius interaction energy. The corresponding quadratic cross-term prediction of the observed I_G achieves the artifact-exposed

$$R^2_{\mathrm{no-fit}} = 0.999585.$$

This is reported as a consistency check of the regime-specific additive organization rather than an independent contribution.

G Out-of-Family and Pretrained-Model Validation

Table 3: External-validation evidence and non-transfer boundaries. Rows are independent statistical units; pooled memory/token rows are not treated as replicates.

Setting	Unit/denominator	Positive evidence	Explicit boundary
Visual writable memory	12 frozen seeds	$E_{\mathrm{SD}} = .00208136$, 12/12 positive; no-fit individual/population $R^2 = .998071/.998089$	Operating-point support/response diagnostic; future accuracy 1.0 (no accuracy gain); no A/u transfer tested
TTT-MLP-125M	12 deterministic non-overlapping held-out windows	P3b centered Pearson/sign $= .740/.747$ (window-cluster CIs $[.657, .820]/[.703, .783]$); P2 redistribution $R^2_{\mathrm{no-fit}} = .952933$	Same checkpoint/focal layer-6; centered relative only; P3a absolute/common-mode $R^2 = -4.430$ fails

G.1 Writable visual memory

Frozen configurations, seed-level rows, and analyses are under `experiments/B-EXTERNAL-001/`.

The visual-memory validation uses CIFAR-10 images, frozen ImageNet-pretrained ResNet18 representations, PCA-reduced 64-dimensional features, and a soft content-addressed writable memory. This setting is outside the original synthetic and recurrent memory families and is evaluated at one operating point.

Read/write definition. There are $N = 20$ unit-normalized targets $p_i \in \mathbb{R}^{64}$, three replay cycles (60 writes), and $\eta = .03$; the present-only calibration selects addressing scale $\beta = 4$. Let $M_t \in \mathbb{R}^{20\times 64}$ contain row-wise writable slots, initially $(M_0)_{i,:} = p_i^\top$. For the frozen augmented-image cue c_t assigned to target j_t,

$$a_t = \mathrm{softmax}_{\mathrm{slots}}(\beta M_t c_t), \qquad r_t = \mathcal{N}(M_t^\top a_t),$$
$$\widetilde{M}_{t+1} = M_t + \eta\,\mathrm{diag}(a_t)(\mathbf{1}z_t^\top - M_t), \quad M_{t+1} = \mathcal{N}_{\mathrm{rows}}(\widetilde{M}_{t+1}).$$

Here $\mathcal{N}(x) = x/\max(\|x\|_2, 10^{-12})$, and $\mathcal{N}_{\mathrm{rows}}$ applies it to each slot. The self arm writes $z_t = r_t$, the oracle writes $z_t = p_{j_t}$, and the frozen arm skips writeback. The exogenous error map instead writes $z_t = \mathcal{N}(p_{j_t} + e_t)$, retaining state-dependent addressing and both normalization steps. Terminal quality is $Q_i = [\mathrm{softmax}_{\mathrm{slots}}(\beta M_T p_i)]_i$. The response differentiates this full map at

$e = 0$, including addressing and normalization. Each seed has 32 held-out error trajectories, with $\|e_t\|_2 = .04$ per write. The endogenous calculation uses the separately recorded self-path residuals $r_t^{\text{self}} - p_{j_t}$; it does not predict those residuals from the initial memory.

The experiment yields

$$E_{\text{SD}} \approx 0.002081, \qquad 95\%\,\text{CI} \approx [0.001830, 0.002364],$$

with $12/12$ positive seeds.

Frozen held-out exogenous prediction achieves no-fit $R^2 = 0.99807$ for individual response and 0.99809 for population redistribution. Conditioning on realized self-path reconstruction errors gives $R^2 = 0.99837$ for endogenous change.

Finite-difference verification gives maximum relative error approximately 1.6×10^{-7}. Aggregate future accuracy remains unchanged; no accuracy gain or A/u transfer is claimed. The 12-seed, single-operating-point experiment supports a continuous support/response diagnostic. It does not establish performance in a capacity-stressed regime with accuracy loss or broad visual-domain transfer.

G.2 Held-out-window replication on the same checkpoint and focal executor

Window selection, execution index, raw window rows, and summaries are under `experiments/B-MODERN-001R/`.

The modern-model validation keeps the same pretrained TTT-MLP-125M checkpoint and focal layer-6 executor. It excludes all original evaluation windows and uses 12 deterministic, mutually non-overlapping held-out windows. Each window contains 64 probe tokens followed by 48 future tokens (three minibatches of 16). Both P2 and P3 use those same future inputs, initial writable state, and horizon; P3 is not a longer-horizon test.

Mapping the focal update to the error coordinate. For each head, the writable state is $\theta = (W_1, b_1, W_2, b_2)$, and its read-only reconstruction is

$$\widehat{r}(\theta, XK) = \text{LN}_{\gamma,\beta}(\text{GELU}_{\tanh}(XKW_1 + b_1)W_2 + b_2)\,.$$

Layer normalization uses the frozen learned affine parameters and variance stabilizer 10^{-6}. The native reconstruction target is $y^{\text{nat}} = XV - XK$. Starting from θ_0 after the 64-token prefix, freeze the base target $y_b^0 = \widehat{r}(\theta_0, XK_b)$ for each future minibatch $b = 0, 1, 2$. The error map replaces the update target with $y_b^0 + e_b$:

$$\theta_{b+1} = \mathcal{U}_b(\theta_b; y_b^0 + e_b, X_b, \Lambda_b), \qquad Q_{\text{final}}(e) = \big(Q_i(\theta_3(e))\big)_{i=1}^{64}.$$

$\mathcal{U}_b$ is the packaged native TTT-MLP minibatch gradient update (`optimized_ttt_mlp_batch`); X_b contains the fixed focal inputs and Λ_b the native head/token learning-rate tensor. The implementation retains its layer-normalization/GELU derivatives, triangular contractions, and final-token weight/bias state updates. It does not replace these native rates with a newly swept scalar step size. Operationally the response is $J_{\text{TTT}} = D_e Q_{\text{final}}(0)$, corresponding to the product ηV in Section 2; its Jacobian–vector product already includes the native writeback rates. No additional η factor is multiplied into the reported TTT prediction. The oracle uses $e = 0$, with the zero-update invariant audited against θ_0. The score Q_i is defined in Appendix A, not a full-model token-loss endpoint.

P2: held-out exogenous perturbations. For each window, 32 independently frozen unit directions $d \in D_B$ are scaled by $a = .05\|y^{\text{nat}}\|_2$. The reported comparison is $aJ_{\text{TTT}}d$ against $Q_{\text{final}}(ad) - Q_{\text{final}}(0)$, and its centered population energy. The amplitude rule is fixed but uses the native target norm; the original P2 is not an absolute target-blind amplitude protocol. It yields $12\times32\times64 = 24{,}576$ individual rows and 384 redistribution rows, with 12 windows as the statistical units.

P3a/P3b: native-target discrepancy. Set $e_b^{\text{nat}} = y_b^{\text{nat}} - y_b^0$. P3a compares the uncentered $J_{\text{TTT}}e^{\text{nat}}$ with $Q_{\text{native}} - Q_{\text{oracle}}$, giving 64 memory/token rows per window. P3b uses exactly these same predictions and outcomes, each centered separately across the 64 probes within its own window. Thus P3b tests relative ordering and centered sign, not recovery of the missing absolute/common-mode component. The native target discrepancy and extracted future inputs are conditional inputs, not quantities forecast from the current memory alone.

Chronology and scope. The original runner caches native/oracle endpoints before response analysis; this is not a claim that numerical predictions were historically serialized first. Predictor construction does not fit to those endpoints. The separate `experiments/B-INTEGRITY-MODERN-001/` fixed-amplitude, ordered-commitment control has its own provenance and does not retroactively alter B-MODERN-001R's inputs or execution order. Twelve selected windows support this scoped replication, not population-wide language, checkpoint, layer, or executor generalization.

Checkpoint provenance matches all $326/326$ tensors exactly, with maximum numerical response-audit error

$$1.07 \times 10^{-8}.$$

The replication recovers prediction of structured/centered future redistribution in P3b, while the P3a absolute/common-mode prediction criterion fails. The result therefore supports held-out-window replication of centered relative structure on this checkpoint and focal executor, not calibrated absolute prediction of all future quality changes. Across the 12 windows, the predicted common shift averages $+0.0307452$, the actual shift averages -0.0140361, and their difference averages -0.0447812, with between-window SD 0.0107885. These shifts are reported diagnostically and are not used to recenter or recalibrate the frozen prediction.

The packaged P2 pooled point metrics are recorded in `experiments/B-MODERN-001R/results/summary.json`. Their row-level inputs are the files `experiments/B-MODERN-001R/results/raw/window_{000..011}_p2.csv` and `experiments/B-MODERN-001R/results/raw/window_{000..011}_p2_directions.csv`.

A separate read-only post-freeze reanalysis of the preserved P3a/P3b row arrays resamples the 12 windows as clusters for 10,000 deterministic bootstrap replicates (seed 612024). The guarded script is `experiments/B-MODERN-001R/src/reanalyze_uncertainty_b_modern_001r.py`, and its completed output is `experiments/B-MODERN-001R/results/uncertainty_reanalysis.json`. This reanalysis covers P3a/P3b only; no P2 window-cluster interval is claimed in the submitted artifact.

For centered P3b, the intervals are $[0.657437, 0.819831]$ for Pearson, $[0.646506, 0.826780]$ for Spearman, $[0.703125, 0.782552]$ for sign accuracy, $[-0.200029, 0.273504]$ for no-fit R^2, and $[0.455011, 0.569020]$ for through-origin slope. Leave-one-window-out ranges are $[0.740057, 0.764205]$ for sign accuracy and $[0.029090, 0.164837]$ for no-fit R^2. These intervals support only the P3b centered-relative rank/sign claim, not calibrated centered magnitude; P3a's absolute/common-mode failure is retained.

H Negative Results and Scope Boundaries

The following negative results define the supported claim boundary.

- **Clean-scale confound (B-ORTHO-001).** The initial intervention failed because writeback scale altered clean dynamics. It was retained as NO-GO rather than repaired post hoc.
- **Operator boundary (B-ARCH-003).** The mismatch gate introduces $\|e\|_2$ at $e = 0$, preventing a global Fréchet response operator and corresponding G at that point. Directional evidence is retained without claiming universality.
- **Strong interaction boundary (B-INTERACT-001).** The hypothesis that allocation strongly modulates the control effect was not supported, and the preregistered extension was stopped.
- **Factorization boundary.** Full A/u factorization is not claimed. External validation supports structured-response replication rather than complete decomposition.
- **Three-corner boundary.** The three-corner result is a post-factorial zero-refit consistency test, not blind prospective discovery or independent factorization evidence.
- **Direction-resampling boundary.** The B-OOS PASS supports direction-averaged susceptibility under independently resampled errors. Its direction-specific residual does not replicate across streams.
- **TTT common-mode boundary.** Within-window rank/sign and relative-response structure replicates on held-out windows of the same checkpoint and focal executor, whereas calibrated centered magnitude and unrestricted absolute/common-mode prediction do not.

I REPRODUCIBILITY AND ARTIFACT DETAILS

The anonymous supplementary ZIP root file `CLAIM_TO_ARTIFACT_INDEX.md` maps each claim to its contract, runner, frozen rows or arrays, and analysis. The provenance classes and omissions are declared in `PROVENANCE_AND_SCOPE.md`; unavailable byte-identical sources or raw tables are not recreated and relabeled as originals.

A separately recorded post-hoc explanatory control, formulated on 2026-09-22, evaluates known trace decay on the 384 individual-memory calibration rows (Appendix C.3). Using recovered packaged scientific functions, it reconstructs all 48 seed–η cells without fitted parameters, reproducing prior reexecuted target and baseline values to within 6×10^{-16} maximum absolute error. The audit is packaged under `post_hoc/decay_weighted_control/`, containing its protocol, configuration, execution scripts, prediction tables, and verification ledgers. Scientific outputs and code remain byte-preserved, while packaging adaptations are tracked via `freeze_identity.json` and `PORTABLE_METADATA_LINEAGE.json` rather than claiming original frozen-manifest bytes. The root `CLAIM_TO_ARTIFACT_INDEX.md` indexes these verification artifacts directly.

Table 4: Exact supplement source-class ledger and key exceptions. Scientific values are not promoted to a stronger source class than the available evidence supports.

Class	Exact coverage and boundary
Local original	B-ORTHO-001, B-DECOUPLED-001, and B-EXTERNAL-001; included runners/outputs are retained without numerical alteration. B-EXTERNAL now includes `RUN_MANIFEST.json` and `audit/script_sha256.txt`; its weight source is `ResNet18_Weights.DEFAULT`, but the weight-file hash is unavailable.
Archival original	B-OOS-DIRECTION-001; raw table/results and manifest hashes are byte-for-byte recovered. The environment-bound `archive/` runner and original result note named by legacy docs are omitted; those legacy paths are omission references, not packaged files. A portable runner is included.
Recovered + fingerprinted	B-ENDOGENOUS-001, B-ALLOC-002, B-CONTROL-001, B-INTERACT-001, and B-THEORY-003; recovered runners/arrays/rows match frozen fingerprints but are not byte-identical originals. Recovered A/u contracts document the included runners/results; B-CONTROL also has a separate deterministic washout audit with raw output, summary, and hashes.
Recovered original + reexecution	B-THEORY-002; historical report/summary, a portable derivative of the recovered original runner, and separately labeled complete reexecution rows are included. Historical raw CSVs remain unavailable. B-ALLOC-002 additionally includes archival nonlinearity/basis diagnostics and a portable original-runner derivative. Source and packaged hashes are recorded in `RECOVERY_FILE_PROVENANCE.json`.
New post-hoc control	B-THEORY-002: separately recorded post-hoc explanatory audit (2026-09-22) evaluating known trace decay across the 384 individual-memory rows; executed via recovered packaged scientific functions against prior reexecuted original-source outputs rather than raw historical bytes.
Reference reconstruction	B-THEORY-001, B-ARCH-002/003, and B-ALLOC-001; contracts, frozen fingerprints, and recovery notes are retained, but original source bytes/raw tables are unavailable.
Hash-verified frozen snapshot	B-MODERN-001R; 144 raw window artifacts, execution indices, and numerical audits are hash-verified for the same checkpoint/focal executor. Its completed uncertainty JSON is a post-freeze, row-based reanalysis (not a model rerun); the historical stop record remains.
Intentional omissions	Third-party weights, tokenizer/data files, large caches, superseded raw artifacts, and the environment-bound archival B-OOS runner; omitted bytes are not recreated or relabeled as originals.
Key numeric exceptions	B-ENDOGENOUS preserves archived CI $[.00432750, .00761454]$ and recovered-summary CI $[.00434354, .00765131]$; the reported gate uses the recovered label. The clean-preactivation diagnostic $L_\phi = .188091$ is distinct from the off-manifold equivariance error 0.182388. B-THEORY-002 finite-difference discrepancies between historical records are retained in the supplementary provenance note. B-THEORY-003 uses summary-exposed $R^2 = .999466$, sign 99.13%, and interaction $R^2 = .999585$.

For B-OOS, `experiments/B-OOS-DIRECTION-001/RUN_MANIFEST.json` stores full digests for the archival runner and 1,152-row raw table. The package includes a portable runner, frozen rows, and analyses, but omits the environment-bound archival copy. *Recovered frozen* denotes preserved information flow and artifacts, not a newly asserted preregistration; raw-table verification passes and the full model was not rerun.

The controlled A/u, visual, and TTT evidence is under `experiments/B-ALLOC-002/`, `experiments/B-CONTROL-001/`, `experiments/B-EXTERNAL-001/`, and `experiments/B-MODERN-001R/`; the last includes deterministic selection, checkpoint provenance, execution indices, 144 raw artifacts, and summaries.

Derivative, present-matching, symmetry, washout, and decision audits accompany their experiments. All figures are generated from or verified against frozen artifacts; no unpublished future result is needed for a main-paper claim.