LOCATING MEMORY IN PREDICTIVE WORLD MODELS: INTERVENTIONAL DIAGNOSTICS BEYOND PROBING

Anonymous authors
Paper under double-blind review

ABSTRACT

Linear probing is the default instrument for asking what a learned representation encodes, but it cannot say *how* a model retains that information. We exhibit three world models — a stateless Joint Embedding Predictive Architecture (JEPA), a GRU and an LSTM — that reach identical probe accuracy (1.000) on the same delayed working-memory task while implementing mechanistically incompatible forms of memory. To separate them we introduce nested *interventional* diagnostics: *stable retention* (SR) asks whether information about a past causal event survives a latent rollout; an *interventional causal separation test*, with a 2×2 extension that toggles recurrent state, asks *where* the memory resides; and the *causal-subspace Lyapunov factor* $\lambda_{\mathbf{v}}(t) = \left\| J_{g_\phi^{(t)}} \mathbf{v} \right\|$ asks *how* the predictor sustains it. Three findings follow. Memory locus is a property of the trained model rather than of the architecture class: GRU and LSTM seeds trained identically span from near-pure state-based to predominantly dynamical solutions, a spread invisible to probe accuracy. A stateless predictor sustains memory by *selective subspace preservation*: at a prediction horizon of $K{=}5$ it contracts sampled latent distances on average ($\hat{\kappa}_\rho \approx 0.36$) while amplifying the causal direction $6.48\times$ relative to a typical direction, a condition strictly weaker than the global norm preservation that unitary and orthogonal recurrent parameterisations enforce. And a controlled ablation that rotates a *trained* encoder's causal weight away from its learned direction, holding predictor, task and training protocol fixed, drives stable retention from 0.47 to 0.07; an orientation-matched control shows the effect follows angular displacement rather than how the weight is spread across coordinates, so encoder geometry gates dynamical memory through alignment with the directions the predictor's dynamics sustain.

1 INTRODUCTION

World models that predict in latent space are increasingly used for planning and representation learning (LeCun, 2022; Assran et al., 2023; Bardes et al., 2024; Hafner et al., 2020), and many of the tasks they are asked to support require remembering an event that is no longer observable. The standard way to ask whether a learned latent "remembers" something is a linear probe (Alain & Bengio, 2016). A probe answers whether a variable is *decodable*. It does not answer whether the model's dynamics *use* that encoding, nor *where* in the model the information is kept. For a world model, these are the questions that matter: a representation from which the past can be decoded, but whose dynamics erase it after a few rollout steps, is of little use for planning.

Probing cannot separate mechanisms that differ. This limitation is not hypothetical. On NeuroGym DelayMatchSampleDistractor1D (Yang et al., 2019), a stateless JEPA-style predictor, a GRU and an LSTM all reach probe accuracy 1.000 (Section 5.1). The JEPA predictor has no hidden state, so whatever it retains must be carried by its latent dynamics; the recurrent models could use their hidden state, their dynamics, or both. Loss-based removal tests do not separate these cases in delay settings either: the causal variable is absent from the prediction targets during the delay, so deleting its latent direction leaves the loss unchanged. Distinguishing the mechanisms requires intervening on the latent state and measuring the downstream effect on what the model retains — a rung-2 question in Pearl's hierarchy (Bareinboim et al., 2022) that no correlational probe reaches.

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.

Dynamical memory in a stateless predictor. Iterating a stateless predictor g_ϕ on its own output defines an autonomous dynamical system $\mathbf{z}_{t+1} = g_\phi(\mathbf{z}_t)$ whose only state is the predicted latent itself. Information about a past event can survive such a rollout only as a persistent divergence between latent trajectories; we call this *dynamical memory*. It is closely related to line-attractor and non-normal accounts of memory in dynamical systems (Seung, 1996; Ganguli et al., 2008; Goldman, 2009; Maheswaranathan et al., 2019), but those accounts are usually studied in networks trained or designed for a memory task. Here the predictor is trained only to forecast the next latents, and whether it acquires such dynamics — and under what encoder conditions — is an empirical question.

What the diagnostics reveal. Applying the suite gives a mechanistic picture with a practical reading. Work on long-range information in recurrent networks has largely pursued *global* constraints: unitary transitions (Arjovsky et al., 2016) and orthogonal parameterisations (Henaff et al., 2016; Vorontsov et al., 2017) keep the transition map approximately norm-preserving. We find that a JEPA-trained predictor meets a far weaker, *local* condition instead: it sustains one causal direction while contracting typical directions. That anisotropic contraction *can* preserve a direction is not surprising; the point is that it *is* what emerges from predictive training without any constraint, that it is measurable, and that it tells a practitioner which design principle (global norm preservation) is unnecessary for retaining a known, low-dimensional causal signal. Whether that direction is available at all is decided upstream, by how the encoder places the causal variable relative to the rest of the latent distribution.

Contributions.

1. **An interventional diagnostic suite for memory in latent predictors** (Section 3, Table 1). Three nested measurements separate *whether*, *where* and *how* a model retains a past causal event. They apply to any latent predictor with an isolable counterfactual event and are informative where probing and loss-based removal are not.
2. **Memory locus is a property of the trained model, not of the architecture class** (Section 5.1). Across identically trained seeds, $\mathrm{gap}(\text{natural})$ spans -0.074 to $+0.703$ for GRU and -0.192 to $+0.841$ for LSTM. The spread survives every exclusion treatment we tried, is invisible to probe accuracy, and is recoverable only by toggling the hidden state.
3. **A measurable mechanism for stateless memory** (Section 5.2). Memory-exhibiting JEPA predictors sustain selectivity $\gg 1$ while contracting sampled distances on average. The *shape* of $\lambda_{\mathbf{v}}(t)$, not its initial magnitude, predicts retention, which accounts for a non-monotonic dependence on the prediction horizon that a capacity-based account does not predict.
4. **Encoder geometry causally gates dynamical memory, through alignment** (Section 5.5). A within-environment manipulation of a trained encoder, holding signal norm, predictor architecture and data fixed, drives retention from 0.47 to 0.07. An orientation-matched control shows the effect tracks angular displacement from the learned direction, not coordinate-level spread.

2 Related Work

Decodability versus use. Decodability does not imply use (Hewitt & Liang, 2019; Belinkov, 2022). Interventional alternatives remove a linear concept and measure behaviour (amnesic probing, nullspace projection and LEACE; Elazar et al., 2021; Ravfogel et al., 2020; Belrose et al., 2023) or patch activations to localise computation (Geiger et al., 2021; Meng et al., 2022). Our causal-separation test belongs to this family with two differences: its outcome is the separation of counterfactual latent *trajectories* rather than task loss, so it stays informative during a delay; and its 2×2 extension crosses the intervention with a hidden-state manipulation, attributing memory to one of two stores rather than only establishing use.

Memory in dynamical systems. Persistent memory in neural dynamics is classically explained by attractors: a line attractor preserves one direction while contracting others (Seung, 1996), and trained RNNs have been reverse-engineered to find such structure (Sussillo & Barak, 2013; Maheswaranathan et al., 2019). Non-normal and feedforward dynamics offer a second route, retaining transient traces

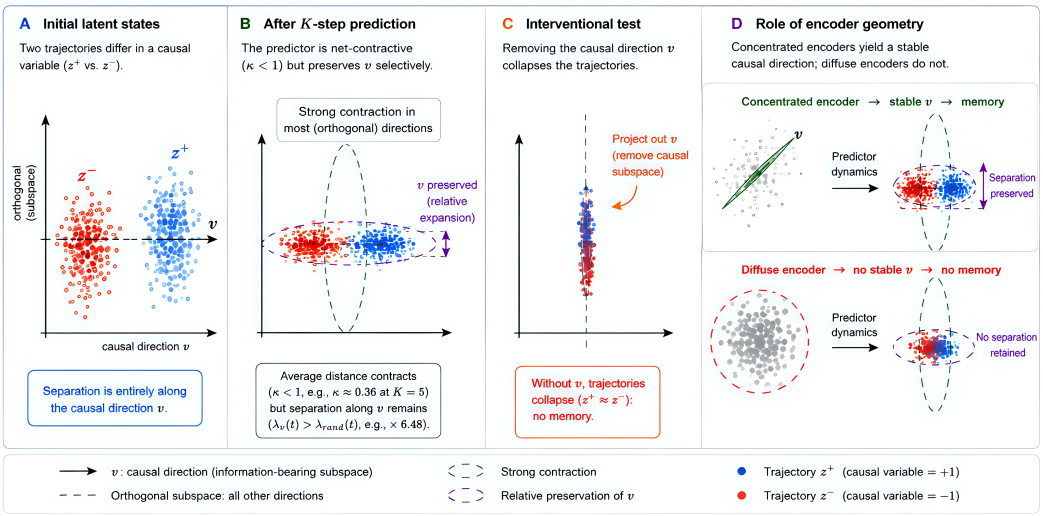


Figure 1: Dynamical memory via selective subspace preservation. **(A)** Two latent trajectories differ along the causal direction $\mathbf{v}$ at the divergence point. **(B)** The predictor contracts sampled latent distances on average yet sustains $\mathbf{v}$, keeping the trajectories apart. **(C)** Interventionally removing $\mathbf{v}$ collapses the separation; removing a random direction does not. **(D)** Whether a sustained causal direction exists is decided upstream, by how the encoder places the causal variable.

without recurrent attractors (Ganguli et al., 2008; Goldman, 2009). Lyapunov analyses characterise information propagation in RNNs *globally*, through leading exponents or full spectra (Engelken et al., 2023; Vogt et al., 2022); a global exponent averages over directions and can report contraction for a map that sustains a task-relevant one. Our $\lambda_{\mathbf{v}}(t)$ is a direction-specific finite-time gain (its logarithm divided by t is the finite-time Lyapunov exponent along $\mathbf{v}$) that isolates exactly that direction. Unitary and orthogonal RNNs (Arjovsky et al., 2016; Henaff et al., 2016; Vorontsov et al., 2017) enforce global norm preservation by construction; our results concern what an unconstrained predictor learns instead.

JEPA and latent world models. JEPA (LeCun, 2022) predicts future latent representations rather than observations; I-JEPA and V-JEPA scale the objective to images and video (Assran et al., 2023; Bardes et al., 2024), and Causal-JEPA introduces object-level latent interventions (Nam et al., 2026). Latent world models for control such as Dreamer carry an explicit recurrent state (Hafner et al., 2020); our GRU and LSTM predictors stand in for that design. All models studied here are *small JEPA-style predictors trained from scratch*; transfer of the diagnostics and the mechanism to large pretrained encoders is untested.

Working memory in neuroscience. Lundqvist et al. (2018) argue that sparse, transient activity can suffice for working memory, matching our *transient* regime. The delay tasks we study (Yang et al., 2019) are established computational analogues of primate working-memory paradigms.

3 Diagnostics

3.1 World model and counterfactual construction

We train a JEPA world model with encoder $f_\theta : \mathcal{O} \to \mathbb{R}^d$ (d=128) and *stateless* predictor $g_\phi : \mathbb{R}^d \to \mathbb{R}^d$, minimising the K-step predictive loss

$$\mathcal{L}_K = \mathbb{E}_t \left[\left\| g_\phi^{(K)}(f_\theta(o_t)) - \mathrm{sg}(f_\theta(o_{t+K})) \right\|^2 \right], \tag{1}$$

Table 1: The diagnostic suite. Each measurement answers one question and is silent on the others, so the three are used together.

Diagnostic	Question	Quantity	Silent on
Stable retention	Does memory exist?	SR, Eq. 2	What carries it
Causal separation	What carries it?	gap, Eq. 4	How it is sustained
2×2 extension	Which store?	$\mathrm{gap}(\text{natural})$, $\mathrm{gap}(\text{zeroed})$	How it is sustained
Lyapunov factor	How is it sustained?	$\lambda_{\mathbf{v}}(t)$, Eq. 5	Functional use

where K is the *prediction horizon* (the number of latent steps the predictor is trained to span, $K \in \{1, 3, 5, 7, 10\}$), $g_\phi^{(K)}$ denotes K iterated applications of g_ϕ, and $\mathrm{sg}(\cdot)$ is stop-gradient. The predictor receives the current latent only and no action, so that dynamical memory is isolated as a property of latent geometry rather than of action-conditioned control (Appendix B). All distances are ℓ_2, matching Eq. 1.

For each environment we construct counterfactual trajectory pairs (τ_+, τ_-) that are identical up to a single causal event and receive identical observations after the *divergence point* t_c, the final timestep of the causal stimulus period. Any later latent divergence is therefore attributable to the causal event by construction (Appendix D.4).

3.2 Stable retention: does memory exist?

We roll the predictor forward from t_c, with $\mathbf{z}_\pm^{t+1} = g_\phi(\mathbf{z}_\pm^t)$, and write $\Delta_t = \left\|\mathbf{z}_+^t - \mathbf{z}_-^t\right\|$. Then stable retention is the time-averaged separation after a burn-in T_s, relative to the separation at the divergence point:

$$\mathrm{SR} = \frac{1}{\Delta_0} \cdot \frac{1}{T - T_s + 1} \sum_{t=T_s}^{T} \Delta_t, \tag{2}$$

with T_s=30, T=60 and memory threshold ε=0.1. $P(\text{memory})$ is the fraction of seeds with SR $> \varepsilon$; Appendix E.5 reports its sensitivity to ε and T_s. For continuous causal variables Δ_0 is averaged over N=100 pairs. Seeds converging to *identity* solutions retain information trivially and are flagged (Appendix D).

3.3 Interventional causal separation: what carries memory?

SR shows that separation persists, not that any particular direction carries it. We estimate the causal direction as the normalised mean counterfactual displacement,

$$\mathbf{v} = \frac{\overline{\mathbf{z}_+ - \mathbf{z}_-}}{\left\|\overline{\mathbf{z}_+ - \mathbf{z}_-}\right\|}, \qquad \overline{\mathbf{z}_+ - \mathbf{z}_-} = \frac{1}{N} \sum_i (\mathbf{z}_+^i - \mathbf{z}_-^i), \tag{3}$$

and remove either $\mathbf{v}$ or a random unit vector $\mathbf{v}_{\text{rand}}$ by orthogonal projection *before* rollout: **Full** (unmodified), **Removed-v** ($\hat{\mathbf{z}}_\pm = \mathbf{z}_\pm - (\mathbf{z}_\pm \cdot \mathbf{v})\mathbf{v}$, then rolled forward) and **Removed-rand** (the same with $\mathbf{v}_{\text{rand}}$, a specificity control). With

$$\mathrm{collapse}(\mathbf{v}) = 1 - \frac{\left\|\hat{\mathbf{z}}_+^{(T_s)} - \hat{\mathbf{z}}_-^{(T_s)}\right\|}{\left\|\mathbf{z}_+^{(T_s)} - \mathbf{z}_-^{(T_s)}\right\|}, \qquad \mathrm{gap} = \mathrm{collapse}(\mathbf{v}) - \mathrm{collapse}(\mathbf{v}_{\text{rand}}), \tag{4}$$

a gap near 1 means that removing $\mathbf{v}$ destroys the separation while removing an arbitrary direction does not, so $\mathbf{v}$ is the dominant carrier of memory. A gap near 0 means $\mathbf{v}$ has no privileged role: either there is no separation to carry, or it is carried elsewhere (the Removed-rand column distinguishes the two). Intermediate values mean the separation partly survives the removal of $\mathbf{v}$, through other latent directions or through a hidden state.

2×2 extension for recurrent architectures. For GRU and LSTM the test runs under two hidden-state conditions: **Active-h** (normal operation) and **Zeroed-h** (h, and for LSTM also c, set to zero

at t_c). The interpretive rule is deliberately asymmetric. A high $\mathrm{gap}(\text{zeroed})$ is *necessary but not sufficient* for dynamical memory, because a state-based model deprived of its state may fall back on its dynamics; only $\mathrm{gap}(\text{natural}) \approx +1$ shows that the dynamics carry the memory in normal operation. This separates *pure dynamical* (both gaps high), *pure state-based* ($\mathrm{gap}(\text{natural}) \approx 0$, $\mathrm{gap}(\text{zeroed}) \approx +1$) and *mixed* (intermediate $\mathrm{gap}(\text{natural})$) solutions.

3.4 Causal-subspace Lyapunov factor: how is memory sustained?

The interventional test localises memory; the Lyapunov factor characterises the dynamics that sustain it:

$$\lambda_{\mathbf{v}}(t) = \left\| J_{g_\phi^{(t)}} \cdot \mathbf{v} \right\|, \qquad \text{selectivity}(t) = \frac{\lambda_{\mathbf{v}}(t)}{\lambda_{\mathbf{v}}^{\text{global}}(t)}, \tag{5}$$

where $J_{g_\phi^{(t)}}$ is the Jacobian of the t-step predictor at a point $\mathbf{z}_0$ from the counterfactual distribution, computed by chained Jacobian-vector products without materialising the Jacobian and averaged over $N{=}100$ starting points, and $\lambda_{\mathbf{v}}^{\text{global}}(t)$ is the same quantity averaged over $N{=}10$ random unit vectors. Selectivity > 1 at large t means the predictor sustains the causal direction better than a typical one.

$$\text{AUC}_{\lambda_{\mathbf{v}}} = \textstyle\sum_{t=0}^{T_s} \lambda_{\mathbf{v}}(t) \tag{6}$$

summarises the profile over the burn-in window.

Four dynamical regimes. Four profile shapes organise the results: **contractive** (monotone decay; no memory), **transient** (early peak then decay; memory bounded by the decay time), **stable** (sustained plateau, selectivity $\gg 1$) and **near-critical** (large initial peak, then rapid collapse; unreliable memory). Regime labels are assigned from profile shape, never from SR magnitude.

3.5 Net contraction, stated precisely

We quantify average contraction as the mean sampled one-step distance ratio under an explicit sampling distribution ρ:

$$\hat{\kappa}_\rho = \mathbb{E}_{(\mathbf{z},\mathbf{z}')\sim\rho}\left[\frac{\|g_\phi(\mathbf{z}) - g_\phi(\mathbf{z}')\|}{\|\mathbf{z} - \mathbf{z}'\|}\right], \tag{7}$$

estimated over $N{=}100$ pairs drawn from the latent sampling distribution used for evaluation in the released code (Appendix D.1). $\hat{\kappa}_\rho < 1$ establishes contraction *on average under* ρ; it is not a Lipschitz bound and is compatible with expansion along selected directions, which is exactly the claim. Indeed the predictor is demonstrably *not* a contraction in the operator-norm sense: at $K{=}5$, $\lambda_{\mathbf{v}}(30) = 1.269 \pm 0.430$, so the 30-step Jacobian has a singular value of at least 1.269, while $\hat{\kappa}_\rho \approx 0.36$. We use "net contraction" only in the sense of Eq. 7.

4 Environments

Four environments span causal-signal structure from a single clean binary direction to an intermittently observable positional signal (full specifications in Appendix C). **NeuroGym DelayMatchSampleDistractor1D** (Yang et al., 2019) presents a sample stimulus, an all-zero delay (2.3 ± 1.9 steps) and a test stimulus; the causal variable is binary stimulus identity and no input is active during the delay. It is the clean positive case and the setting for the architecture comparison. **NeuroGym DelayComparison** uses a continuous magnitude $v_1 \in [10, 34]$ and a fixed 10-step delay, so a single $\mathbf{v}$ only approximates the causal subspace; it measures the cost of that approximation. **BabyAI-Unlock** (Chevalier-Boisvert et al., 2019) requires retrieving a key and unlocking a door; door state is visible only within an egocentric view (in-view run lengths 8.3 ± 11.6 steps) and entangled with agent position. It is the clean negative case. **MiniHack-KeyRoom-S5** (Samvelyan et al., 2021) supplies a binary has-key indicator h to the encoder alongside CNN features c, so the encoder output is *linear* in the causal variable, $\mathbf{z} = W_{\text{visual}} c + w_{\text{key}} h + b$. This gives an analytic handle on encoder geometry, which the ablation of Section 5.5 exploits.

Table 2: 2×2 interventional causal separation on NeuroGym DelayMatchSampleDistractor1D at K=5. All three architectures reach probe accuracy 1.000. gap(natural) is the discriminator of memory locus. *JEPA has no hidden state, so its two gaps coincide by construction. Per-seed values are in Appendices F.3 and F.4; GRU aggregates exclude two expansion-artefact seeds, whose effect is analysed in Appendix F.1.

Model	gap(natural)	Range of gap(natural)	gap(zeroed)	Locus
JEPA	1.000^*	—	1.000^*	Pure dynamical
GRU	$+0.249 \pm 0.249$	-0.074 to $+0.703$	$+1.032 \pm 0.132$	Seed-dep.
LSTM	$+0.275 \pm 0.299$	-0.192 to $+0.841$	$+0.993 \pm 0.024$	Seed-dep.

5 RESULTS

Statistical reporting. Seed-level distributions are heavy-tailed, so claims rest on P(memory) — a threshold count insensitive to one extreme seed — and on $\lambda_{\mathbf{v}}(t)$ profile shape. A single exclusion policy applies uniformly across environments (Appendix A): a seed leaves *gap* averages if collapse($\mathbf{v}$) < -0.5 (*expansion artefact*, a degenerate ratio), and *SR* averages only if it meets the training-instability criterion of Appendix A.1. Where a claim turns on a mean, we report its sensitivity.

5.1 IDENTICAL PROBE ACCURACY, DIFFERENT MEMORY: LOCUS VARIES WITHIN AN ARCHITECTURE

Every seed of both recurrent architectures reaches gap(zeroed) $\approx +1$ (Table 2): deprived of its hidden state, any of these models can fall back on its dynamics. The variation lies entirely in whether $\mathbf{v}$ is *load-bearing when h is present*, and there the same architecture, task and hyperparameters produce solutions from pure state-based (LSTM seed 6, gap(natural) $= +0.001$) to predominantly dynamical (LSTM seed 7, $+0.841$). Seed 7 retains little separation (SR $= 0.004$), which makes its gap ratio poorly conditioned; the spread does not depend on it, since seeds with clear memory already reach $+0.668$ (LSTM seed 0) and $+0.703$ (GRU seed 9). All ten GRU and nine of ten LSTM seeds exceed the SR threshold, so variation in *where* memory lives is not confounded with *whether* it exists.

Sensitivity to exclusions. A claim about *spread* should not depend on which tail is removed. When GRU gap(natural) is recomputed from all ten raw seeds (Appendix F.1), the *mean* is threshold-sensitive and flips sign without exclusions (-0.084 ± 0.813), driven by one seed (-2.330) roughly 3σ from the rest. The *spread* is not: seeds range from near-zero or negative to strongly positive under every treatment, including the exclusion-free median ($+0.103$, MAD 0.261).

Takeaway. Architecture class constrains which stores are *available* — JEPA sits at the dynamical extreme because it has no alternative — but not which are *used*. Describing a trained recurrent model as "storing" a variable in its hidden state is a claim about an individual model that probing cannot support and the 2×2 test can check ($\lambda_{\mathbf{v}}(t)$ profiles for all three models are in Figure 3).

5.2 HOW A STATELESS PREDICTOR SUSTAINS MEMORY: $\lambda_{\mathbf{v}}(t)$ SHAPE PREDICTS RETENTION

On DelayMatchSampleDistractor1D, JEPA retention is *non-monotonic* in the prediction horizon: P(memory) is 0%, 70%, 70%, 20% and 70% at $K = 1, 3, 5, 7, 10$ (Figure 2, Appendix E). A capacity- or horizon-based account predicts improvement with K; the K=7 dip between two 70% conditions is instead explained by the shape of $\lambda_{\mathbf{v}}(t)$ (Figure 2). K=1 decays monotonically to near-zero; K=3 peaks near t=5 and then decays; K=5 sustains a plateau with selectivity 6.48 at t=30 while $\hat{\kappa}_\rho \approx 0.36$; K=7 has a *larger* initial peak than K=5 but collapses after $t \approx 10$; and K=10 sustains a plateau at the highest selectivity (13.38).

Why a decaying profile can still suffice. The transient regime (K=3: decaying $\lambda_{\mathbf{v}}(t)$, yet 70% P(memory)) is consistent once sufficiency is read as a *boundary condition*: the causal signal need

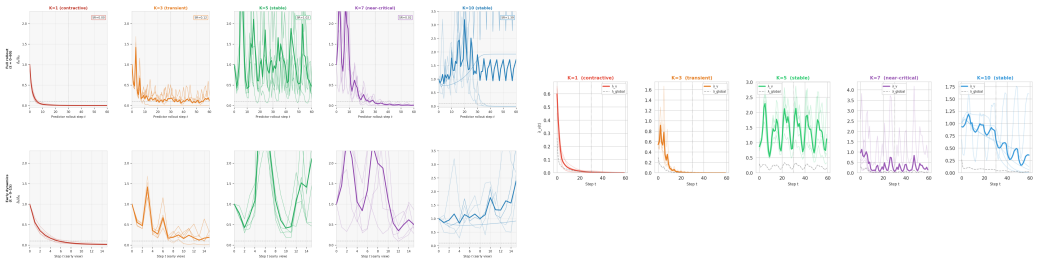


Figure 2: NeuroGym DelayMatchSampleDistractor1D across horizons K. **Left:** stable-retention trajectories, mean Δ_t/Δ_0 over 10 seeds (dashed: $\varepsilon = 0.1$; vertical: $T_s = 30$). **Right:** $\lambda_{\mathbf{v}}(t)$ (solid) and random-direction reference (dashed). Stable regimes sustain the causal direction under net-contractive average dynamics; the near-critical regime (K=7) collapses after a larger early peak.

only survive until the delay ends and observations resupply it, so a peak outlasting the task's mean delay (2.3 steps) suffices even though the profile later decays. This reading makes a falsifiable prediction, which holds: the same K=3 configuration fails on DelayComparison's fixed 10-step delay (10% P(memory), Appendix E.1).

Takeaway. Initial amplitude does not predict retention; profile shape does. The AUC over $[0, T_s]$ separates stable horizons from the rest by a factor of two ($\{39.5, 79.1\}$ versus ≤ 19.8; Table 10), giving a reward-free criterion for choosing K from a coarse grid. The combination of $\hat{\kappa}_\rho < 1$ with a sustained plateau is the signature of an approximately marginal direction inside an otherwise contracting flow — the discrete-time analogue of a line attractor (Seung, 1996) — learned here from a pure prediction objective.

5.3 Across environments: memory requires a stable causal direction

Gap falls from 1.000 on the binary NeuroGym task to 0.679 on continuous DelayComparison and -0.054 on BabyAI (Table 8, Appendix E.2), with probe accuracy 1.000, 0.963 and 0.668. On BabyAI neither $\mathbf{v}$ nor a random direction affects separation, which distinguishes "no memory to carry" from "memory carried elsewhere"; with $\lambda_{\mathbf{v}}(t)$ collapsing immediately at every K (Appendix G) and a predictor identical to the one that succeeds on NeuroGym, the bottleneck is upstream of the predictor. Read across environments, however, this is an association confounded with signal type, observability and delay structure. The next two subsections remove the confounds: Section 5.4 for the one-direction approximation and Section 5.5 for encoder geometry.

5.4 How many dimensions carry a continuous causal signal?

For a binary cause a single direction is exact; for a continuous one it is an approximation. We test whether the DelayComparison shortfall is that approximation or absent memory by replacing $\mathbf{v}$ with the top-k right singular vectors V_k of the counterfactual displacement matrix, projected out together, against a *dimension-matched* random subspace (removing k arbitrary directions destroys more variance than removing one).

At k=1 the extension reproduces the single-direction test (collapse 0.699 over all ten seeds, against 0.657 over the eight seeds retained in Table 8). Two directions recover 0.937 and four saturate at 0.990, while the seed spread narrows from ± 0.222 to ± 0.009 ($\rho(k, \text{gap}) = +0.518$, $p < 10^{-4}$). The shortfall is therefore the one-dimensional approximation, not absent memory.

Specificity and off-manifold effects. A larger non-causal control, removing the leading ambient principal component, expands rather than collapses separation (collapse -1.31, 9 of 10 seeds), and although projecting out $\mathbf{v}$ displaces the latent by 2.27 ± 0.61 ambient standard deviations, the predictor's one-step displacement is unchanged (ratio 1.011). On a continuous variable $\mathbf{v}$ is partly aligned with the leading ambient direction ($\cos = 0.803$), so magnitude and content cannot be fully separated on this task (Appendix E.4).

Table 3: Causal-subspace dimension on DelayComparison (K=5, all 10 seeds). Collapse(V_k) is the fraction of trajectory separation destroyed by removing the top-k causal subspace; the control removes k random orthonormal directions. EVR is the cumulative share of counterfactual displacement variance captured by V_k.

k	Collapse(V_k)	Collapse(random$_k$)	Cumulative EVR
1	0.699 ± 0.222	$+0.008 \pm 0.044$	0.948
2	0.937 ± 0.087	$+0.007 \pm 0.091$	0.996
3	0.957 ± 0.083	-0.053 ± 0.145	1.000
4	0.990 ± 0.009	-0.048 ± 0.147	1.000
8	0.992 ± 0.010	-0.098 ± 0.203	1.000

Table 4: Encoder-rotation ablation on MiniHack-KeyRoom-S5 (K=5, 5 seeds, fresh predictor per variant). Both arms displace w_{key} by the same angle from its trained direction; only the spread arm redistributes it across a basis. At α=0 the two arms are the same model.

		Stable retention		$\lambda_{\mathbf{v}}$ AUC		
α	cos	spread	rotated	spread	rotated	gap($\mathbf{v}$)
0.0	1.000	0.467 ± 0.195	0.472 ± 0.203	16.6 ± 4.3	16.5 ± 4.4	1.00
0.4	0.775	0.322 ± 0.140	0.352 ± 0.174	11.4 ± 3.9	12.3 ± 4.8	0.95
0.8	0.447	0.232 ± 0.072	0.208 ± 0.054	8.5 ± 1.4	7.8 ± 1.6	1.00
1.0	0.000	0.071 ± 0.054	0.039 ± 0.027	3.9 ± 1.6	2.8 ± 0.8	1.12

5.5 Manipulating the encoder: memory follows alignment, not spread

To test the encoder-geometry claim causally, we manipulate a trained encoder within one environment. On MiniHack the encoder is linear in h, so a single column w_{key} carries the causal variable and can be rewritten without touching anything else.

Design. For each of 5 seeds we train a JEPA world model to convergence (K=5), freeze the encoder, and construct variants indexed by α that redistribute w_{key} across a 32-dimensional orthonormal basis by an energy-preserving rotation: α=0 is the trained encoder and α=1 spreads the weight evenly over 31 further directions. $\|w_{\text{key}}\|$ is unchanged, so the counterfactual signal magnitude is fixed and only its placement varies. A fresh predictor is trained on each frozen variant. Counterfactual pairs differ in h alone, so the displacement at the divergence point is exactly $w_{\text{key}}(\alpha)$ (Appendix E.3).

An orientation-matched control. Spreading energy across a basis also *rotates* w_{key} away from its trained direction (cosine $\sqrt{1-\alpha}$). A second arm therefore rotates w_{key} by the *same angle* about a random axis: identical norm and cosine, no redistribution. Both arms share the encoder, predictor initialisation and training stream at every level, so the comparison is paired at the level of each (seed, α) cell. At α=0 both arms are the unmodified encoder, a null condition that measures the pipeline's own run-to-run noise.

Result. Retention falls monotonically with α in *both* arms: SR from 0.47 to 0.07 and $\lambda_{\mathbf{v}}$-AUC from 16.6 to 3.9 under spreading, and from 0.47 to 0.04 and 16.5 to 2.8 under orientation-matched rotation (Spearman $\rho(\alpha, \text{SR}) = -0.760$ and -0.776; $\rho(\alpha, \text{AUC}) = -0.830$ and -0.861). We detect no difference between the arms: the median paired SR difference is -0.001, smaller than the 0.007 run-to-run discrepancy at the null level, and a paired Wilcoxon test gives $p = 0.90$. We do not claim formal equivalence at five seeds. Encoder geometry therefore moves dynamical memory causally, within a single environment and with signal norm, predictor architecture and data held fixed; what matters is *angular displacement from the learned direction*, not how the weight is distributed across coordinates.

Where the memory goes. gap($\mathbf{v}$) stays near 1 at every level (Table 4), so $\mathbf{v}$ remains the dominant carrier throughout: rotation does not relocate the memory to another latent direction, it stops the dynamics sustaining it. The natural reading is that the predictor supports only the subspace its own rollout distribution occupies, and encoder geometry matters because it decides whether the causal variable lands inside it. This is consistent with the ambient rollout distribution being effectively rank one on DelayComparison (Appendix E.4), and it accounts for the cross-environment pattern of Section 5.3 without appealing to coordinate-level concentration. We mark it as an interpretation rather than a tested claim: establishing it would require manipulating the overlap between w_{key} and the sustained subspace directly.

A diagnostic dissociation. gap($\mathbf{v}$) does not track α while SR and $\lambda_{\mathbf{v}}$ do: the displacement at t_c is exactly $w_{\text{key}}(\alpha)$ and $\mathbf{v}$ is re-estimated at every level, so removing it collapses separation almost by construction. One manipulation moving two diagnostics and not the third is direct evidence that the suite is non-redundant.

5.6 Control ablations

Four further ablations on MiniHack (5 seeds each; Appendix F.2) bound the claim from both sides. Resetting the predictor abolishes memory at every horizon (SR $= 0.000$), so dynamical memory lives in *learned* predictor weights: a favourably placed causal direction is necessary (Section 5.5) but not sufficient. Freezing the encoder leaves SR comparable to baseline at $K \in \{5, 10\}$ (5/5 seeds), so encoder geometry at the freeze point already suffices for the predictor to learn selective preservation. Linear predictors never reach bounded selective preservation: they collapse to identity at $K{=}1$ (1.037 ± 0.073) or expand unstably at $K \geq 5$ (1.636 ± 0.744 at $K{=}5$). Whether this is a limitation of linear maps or of optimisation under the JEPA loss is open.

6 Discussion

What a practitioner can do with the suite. The diagnostics support *auditing* (checking, for a given trained model, a claim that it stores a variable in its hidden state), *reward-free model selection* ($\lambda_{\mathbf{v}}$-AUC separated stable from unstable horizons by a factor of two from a coarse grid), and *localising failure* (high probe accuracy with gap ≈ 0 and a collapsing $\lambda_{\mathbf{v}}(t)$, as on BabyAI, implicates the encoder, so adding predictor capacity is the wrong fix).

A reliability trade-off against recurrence. All ten GRU and nine of ten LSTM seeds exceed the SR threshold, against 70% for JEPA at $K{=}5$ on the same task. In this small, single-task sample an explicit hidden state buys a more reliable fallback at the cost of parameters and state; dynamical memory is parameter-free but contingent on how the encoder places the causal variable, a contingency Section 5.5 shows is causal. Encouraging a known feature to align with the predictor's sustained subspace could therefore substitute for recurrent state; we have not trained such an objective.

Limitations. $\mathbf{v}$ is exact only for binary causes; the k-dimensional extension recovers the continuous case but is evaluated on one task, with no rule for choosing k. The encoder ablation varies the angle of w_{key}, not the dimensionality of the causal subspace, and the alignment account awaits a direct test. Projection moves the latent about two ambient standard deviations off-manifold, which we bound rather than eliminate. The architecture comparison rests on 18 valid runs at one task and one K; all metrics use ℓ_2; the predictor is not action-conditioned, so nothing uses the retained separation for control; and all models are small and trained from scratch.

7 Conclusion

Probe accuracy at ceiling says that a representation carries a variable, not how the model keeps it. Nested interventional diagnostics show that memory locus is selected during training rather than fixed by architecture, that a stateless predictor retains a causal signal by sustaining a single direction inside a contracting flow, and that the encoder gates this memory through alignment rather than spread. An open question is when predictive training places causal variables where the predictor can sustain them.

AI USE STATEMENT

In this work, we used generative AI tools only for tasks with recommended disclosure: drafting and editing text, improving the paper's structure and readability, identifying and summarising related literature, and preparing the conceptual schematic in Figure 1, which illustrates the mechanism and contains no experimental data. We did not use generative AI tools to design or run experiments, to implement methods or diagnostics, or to generate, clean or process data; all other figures are direct plots of experimental output from the released code. We have reviewed all AI-assisted work: every AI-edited passage was checked by the authors against the experimental results it reports. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This research involves no human subjects and no sensitive or personally identifiable data. All experiments use three publicly available simulated benchmarks, namely NeuroGym (Yang et al., 2019), the BabyAI platform (Chevalier-Boisvert et al., 2019) and the MiniHack sandbox (Samvelyan et al., 2021). We release no pretrained models or datasets with meaningful dual-use potential and foresee no direct path to negative societal impact.

REPRODUCIBILITY STATEMENT

Architectures, optimiser settings and training schedules are in Appendix B; counterfactual construction in Appendices D.4 and E.3; metric definitions, including $\hat{\kappa}_\rho$ and its sampling distribution, in Appendices D and D.1; the $\lambda_{\mathbf{v}}(t)$ implementation in Appendix D.2; and the 2×2 protocol in Appendix D.3. One exclusion policy, applied uniformly, is stated in Appendix A, with sensitivity analyses in Appendices F.1 and E.5. Main results use 10 seeds, ablations 5, and the encoder ablation 5 seeds per level in each of two arms. Per-seed values for every aggregate are in Appendix F. Code for all experiments and diagnostics is included as supplementary material.

ACKNOWLEDGMENTS

[Removed for anonymous submission.]

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv:1610.01644*, 2016.

Martin Arjovsky, Amar Shah, and Yoshua Bengio. Unitary evolution recurrent neural networks. In *ICML*, 2016.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *CVPR*, 2023.

Adrien Bardes, Quentin Garrido, Jean Ponce, et al. Revisiting feature prediction for learning visual representations from video. *arXiv:2404.08471*, 2024.

Elias Bareinboim, Juan D. Correa, Duligur Ibeling, and Thomas Icard. On Pearl's hierarchy and the foundations of causal inference. In *Probabilistic and Causal Inference: The Works of Judea Pearl*. ACM, 2022.

Yonatan Belinkov. Probing classifiers: promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In *NeurIPS*, 2023.

Maxime Chevalier-Boisvert, Dzmitry Bahdanau, Salem Lahlou, et al. BabyAI: A platform to study the sample efficiency of grounded language learning. In *ICLR*, 2019.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, et al. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *EMNLP*, 2014.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the ACL*, 9:160–175, 2021.

Rainer Engelken, Fred Wolf, and L. F. Abbott. Lyapunov spectra of chaotic recurrent neural networks. *Physical Review Research*, 5:043044, 2023.

Surya Ganguli, Dongsung Huh, and Haim Sompolinsky. Memory traces in dynamical systems. *PNAS*, 105(48):18970–18975, 2008.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *NeurIPS*, 2021.

Mark S. Goldman. Memory without feedback in a neural network. *Neuron*, 61(4):621–634, 2009.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *ICLR*, 2020.

Mikael Henaff, Arthur Szlam, and Yann LeCun. Recurrent orthogonal networks and long-memory tasks. In *ICML*, 2016.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *EMNLP*, 2019.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

Yann LeCun. A path towards autonomous machine intelligence. OpenReview preprint, 2022.

Mikael Lundqvist, Pawel Herman, and Earl K. Miller. Working memory: delay activity, yes! persistent activity? maybe not. *Journal of Neuroscience*, 38(32):7013–7019, 2018.

Niru Maheswaranathan, Alex H. Williams, Matthew D. Golub, Surya Ganguli, and David Sussillo. Reverse engineering recurrent networks for sentiment classification reveals line attractor dynamics. In *NeurIPS*, 2019.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *NeurIPS*, 2022.

H. Nam et al. Causal-JEPA: Learning world models through object-level latent interventions. *arXiv:2602.11389*, 2026.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *ACL*, 2020.

Mikayel Samvelyan, Robert Kirk, Vitaly Kurin, et al. MiniHack the planet: A sandbox for open-ended reinforcement learning research. In *NeurIPS Datasets and Benchmarks Track*, 2021.

H. Sebastian Seung. How the brain keeps the eyes still. *PNAS*, 93(23):13339–13344, 1996.

David Sussillo and Omri Barak. Opening the black box: Low-dimensional dynamics in high-dimensional recurrent neural networks. *Neural Computation*, 25(3):626–649, 2013.

Ryan Vogt, Maximilian Puelma Touzel, Eli Shlizerman, and Guillaume Lajoie. On Lyapunov exponents for RNNs: Understanding information propagation using dynamical systems tools. *Frontiers in Applied Mathematics and Statistics*, 8, 2022.

Eugene Vorontsov, Chiheb Trabelsi, Samuel Kadoury, and Chris Pal. On orthogonality and learning recurrent networks with long term dependencies. In *ICML*, 2017.

Guangyu Robert Yang, Madhura R. Joglekar, H. Francis Song, William T. Newsome, and Xiao-Jing Wang. Task representations in neural networks trained to perform many cognitive tasks. *Nature Neuroscience*, 22(2):297–306, 2019.

A Exclusion Policy

One policy applies uniformly across environments and architectures. A seed is excluded from *gap* averages if collapse($\mathbf{v}$) < -0.5 under the condition being averaged (*expansion artefact*): such seeds sit near a stability boundary where the denominator of Eq. 4 is degenerate and the ratio is not interpretable as a collapse fraction. A seed is excluded from *SR* averages only if it meets the training-instability criterion of Appendix A.1. No other exclusions are applied. Raw values for every excluded seed appear in the per-seed tables, and the claim most sensitive to the policy (GRU locus spread) has a full sensitivity analysis (Appendix F.1).

The instability criterion is met only by BabyAI seed 2042 at K=10. NeuroGym K=10 shows comparable seed-level dispersion (Table 6), but no seed there meets the criterion, so none is excluded; we flag the dispersion and read that condition through $P(\text{memory})$ and $\lambda_{\mathbf{v}}(t)$ shape rather than the SR mean.

A.1 Training-Instability Criterion

A seed is flagged when its training loss spikes by more than an order of magnitude after step 1000 *and* the resulting one-step gain exceeds 2.0, indicating gradient explosion rather than a learned solution. BabyAI seed 2042 at K=10 meets both (loss spike to 5.05 at step 1500, $\Delta_0 = 7.73$, SR $= 2.17$, gain 6.81); excluding it gives $P(\text{memory}) = 0$ for all BabyAI conditions.

B Implementation Details

MiniHack encoder. Two-layer CNN (Conv2d 1→32→64, kernel 3, padding 1, ReLU), flatten, Linear $(64 \times 21 \times 79 + 1) \to 128$. The $+1$ feature is the binary has-key indicator, concatenated before the final projection, so the encoder output is linear in the causal variable.

NeuroGym / DelayComparison encoder. Linear(obs_dim, 256), ReLU, Linear(256, 128). LayerNorm on the encoder output in recurrent models.

BabyAI encoder. Two-layer CNN (Conv2d 3→32→64, kernel 3, padding 1, ReLU), flatten, Linear($64 \times 7 \times 7$, 128).

JEPA predictor (all environments). Linear(128, 256), ReLU, Linear(256, 128); Xavier uniform initialisation, zero biases.

GRU / LSTM predictors. Single recurrent layer, hidden dimension 256, linear projection to 128.

Training. Adam; learning rate 10^{-3} (MiniHack, BabyAI) or 3×10^{-4} (NeuroGym, DelayComparison, all recurrent models); gradient clipping at norm 1.0; 2000 steps per condition; single T4 GPU.

Action conditioning. g_ϕ receives only $\mathbf{z}_t$. In action-conditional environments the predictor implicitly marginalises over a fixed behavioural policy. This isolates dynamical memory as a property of latent geometry, and is also why the diagnostics are not connected to control performance.

C Environment Specifications

C.1 NeuroGym Tasks

DelayMatchSampleDistractor1D: sample period 7.9 ± 0.2 steps; all-zero delay 2.3 ± 1.9 steps; binary stimulus identity as causal variable. DelayComparison: continuous magnitude $v_1 \in [10, 34]$, 5-step stimulus, fixed 10-step all-zero delay; counterfactual pairs use maximally separated stimuli ($v_1^+ = 34$, $v_1^- = 10$).

C.2 MiniHack-KeyRoom-S5

A 5×5 room in which the agent retrieves a key and unlocks a door. Observations are a 21×79 symbolic glyph map plus inventory strings, from which the has-key indicator is extracted. For the baseline and control ablations, two environment instances share an identical random action prefix up

to the key-pickup step; in env_+ the pickup proceeds and in env_- it is replaced by a no-op, and both receive identical suffix actions for 40 steps. The encoder ablation uses a different, tighter construction (Appendix E.3).

C.3 BabyAI Signal Structure

Three properties impede a stable causal direction: door state is observable only within the egocentric view; in-view run lengths are highly variable (CV $= 1.39$); and door state is entangled with agent position in the egocentric frame. The interventional test (gap $= -0.054$) and the immediately contractive $\lambda_{\mathbf{v}}(t)$ are consistent with this diagnosis, and the predictor is identical to the one that succeeds on NeuroGym.

Table 5: Causal-signal observability statistics. CV $= \sigma/\mu$; K^* approximates the horizon needed to span the delay ($\mu_{\text{delay}} + \sigma_{\text{delay}}$, rounded). $P(\text{memory})$ is reported at $K{=}5$ in every row.

Environment	Signal type	Run $\mu \pm \sigma$	Delay $\mu \pm \sigma$	CV(delay)	K^*	P(mem)
NeuroGym	Binary	7.9 ± 0.2	2.3 ± 1.9	0.834	9	0.70
DelayComp	Continuous	5.0 ± 0.0	10.0 ± 0.0	0.000	10	0.80
BabyAI	Intermittent	8.3 ± 11.6	3.9 ± 4.4	1.122	11	0.00

D Metric Definitions and Estimation

$P(\textbf{memory})$**.** Fraction of seeds with SR $> \varepsilon = 0.1$ after burn-in $T_s = 30$.

Identity regime. Flagged when training loss falls below 10^{-4} by step 100 *and* one-step gain lies in $[0.95, 1.05]$.

D.1 Estimating $\hat{\kappa}_\rho$

$\hat{\kappa}_\rho$ (Eq. 7) is estimated over $N{=}100$ pairs drawn from the latent sampling distribution used for evaluation in the released code. The estimate is distribution-dependent, so we name ρ wherever it appears and treat it as a sampled average, never as a Lipschitz or operator-norm bound (Section 3.5).

D.2 $\lambda_{\mathbf{v}}(t)$ Implementation

Chained forward-mode automatic differentiation (`torch.autograd.functional.jvp`); each step computes $J_{g_\phi}(\mathbf{z}^{(t)}) \cdot \mathbf{v}^{(t)}$ in $O(d \cdot h)$ without materialising the Jacobian. The tangent is propagated un-normalised, so $\lambda_{\mathbf{v}}(t) = \left\|\mathbf{v}^{(t)}\right\|$ is the cumulative gain and $t^{-1} \log \lambda_{\mathbf{v}}(t)$ the finite-time Lyapunov exponent along $\mathbf{v}$. Averaged over $N{=}100$ starting points; $\lambda_{\mathbf{v}}^{\text{global}}(t)$ over $N{=}10$ random unit vectors.

D.3 2×2 Protocol

$N = 200$ counterfactual pairs for $\mathbf{v}$ estimation, seeds offset by 500,000 from training; $N_{\text{test}} = 50$ pairs for removal; collapse evaluated at $T_s = 30$. Zeroed-h: for GRU, h set to zero; for LSTM, both h_t and c_t.

D.4 Divergence Point

t_c is the final timestep of the causal stimulus period. SR rollout and interventional removal both initialise from latents at t_c. For DelayComparison, $\mathbf{v}$ is estimated from latents at the end of the delay ($t = t_c +$ delay_len), where the continuous signal is most concentrated after propagation.

D.5 K Selection via $\lambda_{\mathbf{v}}(t)$ Shape

Stable horizons have substantially higher $\lambda_{\mathbf{v}}$-AUC over $[0, T_s]$ than contractive or near-critical ones (Table 10). As a heuristic, K should exceed $\mu_{\text{delay}} + \sigma_{\text{delay}}$ but not equal the delay length exactly,

which produces over-compression (DelayComparison $K{=}10$, Table 7). A principled characterisation of optimal K remains open.

E ADDITIONAL RESULTS

Table 6: JEPA stable retention on NeuroGym DelayMatchSampleDistractor1D (10 seeds). Regimes are assigned from $\lambda_{\mathbf{v}}(t)$ shape (Table 10), not from SR. P(Mem) is robust to the heavy tails that inflate the $K{=}10$ standard deviation and is the statistic the regime claims rely on.

K	SR (mean $\pm$ sd)	P(Mem)	Regime
1	0.000 ± 0.000	0%	Contractive
3	0.230 ± 0.197	70%	Transient
5	0.818 ± 0.716	70%	Stable
7	0.157 ± 0.300	20%	Near-critical
10	$1.116 \pm 2.251^{\dagger}$	70%	Stable

†Dispersion exceeds the mean because of a single seed near a stability boundary that does not meet the instability criterion (Appendix A); $K{=}10$ is read through P(Mem) and $\lambda_{\mathbf{v}}(t)$ shape.

E.1 DELAYCOMPARISON SR

Table 7: DelayComparison stable retention (SR averaged over 100 pairs, 10 seeds). At $K{=}10$ the horizon exactly matches the delay length, forcing the predictor to span the gap in one application (over-compression).

K	SR	P(Mem)	Regime
1	0.000 ± 0.000	0%	Contractive
3	0.058 ± 0.055	10%	Transient
5	0.400 ± 0.286	80%	Stable
7	0.732 ± 0.412	90%	Stable
10	0.005 ± 0.006	0%	Contractive

E.2 CROSS-ENVIRONMENT INTERVENTIONAL RESULTS

Table 8: Interventional causal separation for JEPA at $K{=}5$. DelayComparison excludes two expansion-artefact seeds under the policy of Appendix A (8 seeds).

Environment	Probe acc.	Gap	Collapse($\mathbf{v}$)	Collapse(rand)
NGYM-DMSD1D	1.000 ± 0.000	1.000	1.000 ± 0.000	-0.000 ± 0.007
NGYM-DelayComp	0.963 ± 0.057	0.679	0.657 ± 0.263	-0.022 ± 0.031
BabyAI-Unlock	0.668 ± 0.049	-0.054	-0.057 ± 0.194	-0.003 ± 0.014

E.3 COUNTERFACTUAL CONSTRUCTION FOR THE ENCODER ABLATION

Each pair is built from a single rollout, duplicated, with the has-key indicator set in one copy. Observations are otherwise identical, so the latent displacement at the divergence point is exactly $w_{\text{key}}(\alpha)$ on every pair and displacement coherence ($\|\overline{\Delta \mathbf{z}}\|/\overline{\|\Delta \mathbf{z}\|}$) is 1.000 by construction. This isolates the predictor's response to the has-key displacement from positional and visual differences, which the world-level construction of Appendix C.2 also introduces.

E.4 Specificity and Off-Manifold Checks

Magnitude. On DelayComparison the random control removes about 8% of the latent norm and $\mathbf{v}$ removes 58%. A larger non-causal control — the top principal component of the *ambient* post-delay distribution, estimated from ordinary rollouts — removes 36% and yields collapse -1.31: it expands separation on 9 of 10 seeds. Two non-causal interventions differing more than fourfold in removed norm both fail to reproduce the collapse. At k=2, where the controls remove comparable norm, collapse is 0.937 against 0.449: separated, but not cleanly, because $\mathbf{v}$ is partly aligned with the leading ambient direction ($\cos = 0.803$).

Off-manifold displacement. Projecting out $\mathbf{v}$ displaces the latent by 2.27 ± 0.61 ambient standard deviations measured along the displacement direction. We report this rather than a Mahalanobis distance because the ambient post-delay covariance is effectively rank one (participation ratio 1.0 in d=128; condition number $\sim 10^{30}$), so a Mahalanobis estimate is dominated by the regulariser used to invert it (a six-decade ridge sweep moves it from 225 to 27). On the same scale held-out ambient points score 1.32 and unprojected counterfactual latents 1.72. The predictor's one-step displacement is unchanged by projection (ratio 1.011).

A rank-one ambient distribution. The rank-one ambient covariance is itself a signature of selective preservation measured independently of $\lambda_{\mathbf{v}}(t)$: the predictor contracts the ambient distribution onto essentially one direction, which is largely the causal one. A predictor converging to a fixed point would also produce low-rank ambient covariance, so we report the observation as suggestive rather than as independent evidence.

E.5 Sensitivity of $P(\text{memory})$ to ε and T_s

Table 9: $P(\text{memory})$ on DelayComparison at K=5 across thresholds and burn-in points (10 seeds). The reported value (0.80) lies within this range and the regime assignment does not change.

Burn-in T_s	$\varepsilon = 0.05$	$\varepsilon = 0.1$	$\varepsilon = 0.2$
20	1.00	1.00	0.80
30	1.00	1.00	0.80
40	1.00	1.00	0.80

E.6 Lyapunov Factor Summaries

Table 10: Causal-subspace Lyapunov summaries for NeuroGym DelayMatchSampleDistractor1D (5 seeds per K; the SR results of Table 6 use 10 seeds, because the Lyapunov grid was run at lower cost). All K satisfy $\hat{\kappa}_\rho < 1$; at K=5, $\hat{\kappa}_\rho \approx 0.36$ while $\lambda_{\mathbf{v}}(30) = 1.269$.

K	$\lambda_{\mathbf{v}}(30)$	AUC	Selectivity	Pattern
1	0.006 ± 0.004	2.0	1.01	Contractive
3	0.003 ± 0.002	5.4	3.81	Transient
5	1.269 ± 0.430	79.1	6.48	Stable
7	0.050 ± 0.066	19.8	4.71	Near-critical
10	0.672 ± 0.261	39.5	13.38	Stable

E.7 BabyAI SR

Mean $\pm$ sd across 10 seeds: K=1: 0.007 ± 0.005; K=3: 0.022 ± 0.018; K=5: 0.018 ± 0.009; K=7: 0.015 ± 0.020; K=10: 0.227 ± 0.646 before exclusion of seed 2042 (Appendix A.1), $P(\text{memory}) = 0$ after.

F PER-SEED RESULTS

F.1 SENSITIVITY OF THE GRU LOCUS STATISTIC TO EXCLUSIONS

Table 11: Sensitivity of GRU $\mathrm{gap}(\text{natural})$ to exclusion treatment (NeuroGym DelayMatchSampleDistractor1D, $K{=}5$).

Treatment	Statistic	Value
Stated policy ($n{=}8$)	mean $\pm$ sd	$+0.249 \pm 0.249$
No exclusion ($n{=}10$)	mean $\pm$ sd	-0.084 ± 0.813
Seed 3 only excluded ($n{=}9$)	mean $\pm$ sd	$+0.165 \pm 0.334$
Exclusion-free robust ($n{=}10$)	median $\pm$ MAD	$+0.103 \pm 0.261$

F.2 ABLATION RESULTS

Table 12: Ablations on MiniHack for $K \in \{1, 5, 10\}$, 5 seeds each. † identity regime; § expanding instability.

Ablation	$K{=}1$ SR / P(Mem)	$K{=}5$ SR / P(Mem)	$K{=}10$ SR / P(Mem)
Baseline	0.851 ± 0.195 / 5/5	0.481 ± 0.252 / 4/5	0.451 ± 0.300 / 3/5
No predictor (identity)	1.000 ± 0.000 / 5/5†	1.000 ± 0.000 / 5/5†	1.000 ± 0.000 / 5/5†
Frozen encoder	0.853 ± 0.422 / 4/5	0.504 ± 0.151 / 5/5	0.461 ± 0.123 / 5/5
Predictor reset	0.000 ± 0.000 / 0/5	0.000 ± 0.000 / 0/5	0.000 ± 0.000 / 0/5
Linear predictor	1.037 ± 0.073 / 5/5†	1.636 ± 0.744 / 5/5§	1.256 ± 0.214 / 5/5§

F.3 GRU PER-SEED RESULTS

Table 13: Per-seed GRU results (NeuroGym DelayMatchSampleDistractor1D, $K{=}5$). The exclusion criterion is applied to collapse$(\mathbf{v})$ under the natural-h condition, not to the gap shown here; the two differ by collapse$(\mathbf{v}_{\text{rand}})$. Seeds 2 and 3 have collapse$(\mathbf{v}) = -0.514$ and -2.386 (gap -0.503 and -2.330) and are excluded from gap averages; raw values are retained. Seed 2 lies just past the threshold, which is why Appendix F.1 reports every alternative treatment. All ten seeds exceed the SR memory threshold.

Seed	SR	$\mathrm{gap}(\text{natural})$	$\mathrm{gap}(\text{zeroed})$	Status
0	0.699	$+0.460$	$+0.993$	Valid
1	0.550	$+0.077$	$+0.971$	Valid
2	0.497	-0.503	$+0.943$	Excluded[a]
3	0.646	-2.330	$+0.996$	Excluded[b]
4	0.491	$+0.447$	$+1.020$	Valid
5	0.807	-0.005	$+0.999$	Valid
6	0.451	$+0.128$	$+1.354$	Valid
7	0.540	$+0.254$	$+0.951$	Valid
8	0.895	-0.074	$+0.966$	Valid
9	0.857	$+0.703$	$+1.000$	Valid
Mean (8 valid)	0.661	$+0.249 \pm 0.249$	$+1.032 \pm 0.132$	
Median (10 raw)	—	$+0.103 \pm 0.261$ (MAD)	—	

[a] collapse$(\mathbf{v}) = -0.514$; [b] collapse$(\mathbf{v}) = -2.386$ (natural-h condition).

F.4 LSTM PER-SEED RESULTS

Table 14: Per-seed LSTM results (NeuroGym DelayMatchSampleDistractor1D, $K{=}5$). Locus labels follow Section 3.3; "Pred." abbreviates predominantly. ‡Seed 7 retains essentially no separation (SR 0.004), so its gap ratio has a near-zero denominator; its label should be read with that caveat.

Seed	SR	gap(natural)	gap(zeroed)	Locus
0	0.319	$+0.668$	$+0.999$	Mixed
1	0.586	-0.192	$+1.004$	Pred. state-based
2	0.381	$+0.331$	$+1.024$	Mixed
3	0.397	$+0.323$	$+0.930$	Mixed
4	0.437	$+0.065$	$+1.000$	Pred. state-based
5	0.831	$+0.159$	$+1.000$	Pred. state-based
6	0.796	$+0.001$	$+1.000$	Pure state-based
7	0.004	$+0.841$	$+0.999$	Pred. dynamical‡
8	0.743	$+0.457$	$+0.979$	Mixed
9	0.832	$+0.096$	$+0.995$	Pred. state-based
Mean $\pm$ sd	0.533 ± 0.259	$+0.275 \pm 0.299$	$+0.993 \pm 0.024$	

Every seed reaches $\mathrm{gap(zeroed)} \approx +1$; the variation lies in whether $\mathbf{v}$ is load-bearing when h is present. Excluding seed 7 leaves the range -0.192 to $+0.668$, so the spread claim does not depend on it.

F.5 FROZEN-ENCODER RESULTS (MINIHACK)

Table 15: Frozen-encoder ablation, per seed.

Seed	$K{=}1$ SR	$K{=}5$ SR	$K{=}10$ SR
42	1.1472	0.2787	0.4149
1042	1.0117	0.5253	0.6066
2042	0.0251	0.7049	0.4795
3042	0.9951	0.4689	0.5365
4042	1.0846	0.5448	0.2659
Mean $\pm$ sd	0.853 ± 0.422	0.504 ± 0.151	0.461 ± 0.123
P(Mem)	4/5	5/5	5/5

F.6 LINEAR PREDICTOR RESULTS (MINIHACK)

Table 16: Linear predictor ablation, per seed.

Seed	$K{=}1$ SR	$K{=}5$ SR	$K{=}10$ SR
42	0.9973	0.7207	1.2872
1042	1.1853	1.2270	1.4150
2042	0.9995	2.5261	0.8802
3042	1.0000	2.5074	1.4150
4042	1.0006	1.1965	1.2802
Mean $\pm$ sd	1.037 ± 0.073	1.636 ± 0.744	1.256 ± 0.214

G SUPPLEMENTARY FIGURES

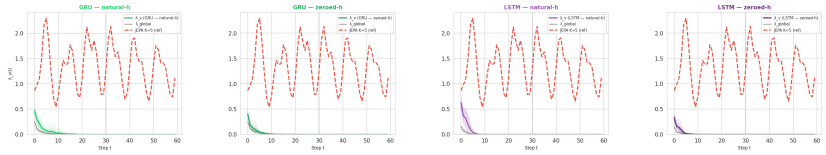


Figure 3: Causal-subspace dynamics on NeuroGym DelayMatchSampleDistractor1D for three architectures with identical probe accuracy (1.000). The profiles differ, but only the 2×2 intervention attributes the difference to a store.

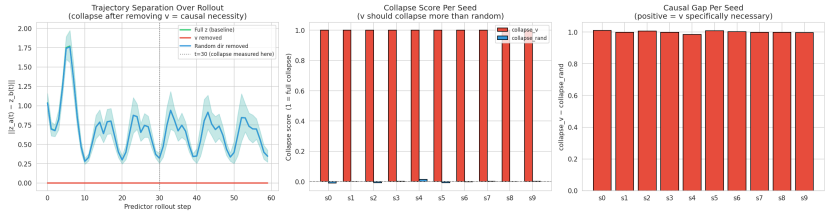


Figure 4: Interventional causal separation on NeuroGym at $K=5$. Removing $\mathbf{v}$ collapses trajectory separation; removing a random direction has no effect.

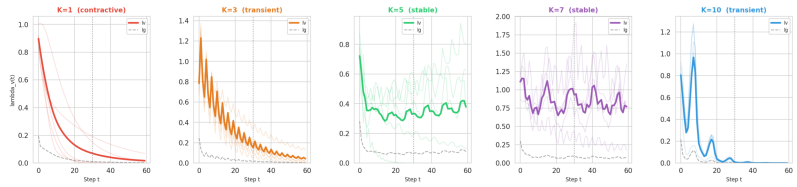


Figure 5: Causal-subspace Lyapunov factor on DelayComparison. Solid: $\lambda_{\mathbf{v}}(t)$; dashed: $\lambda_{\mathbf{v}}^{\text{global}}(t)$. $K \in \{5, 7\}$ sustain elevated but noisy amplification; $K=10$ decays rapidly, consistent with over-compression at the exact delay length.

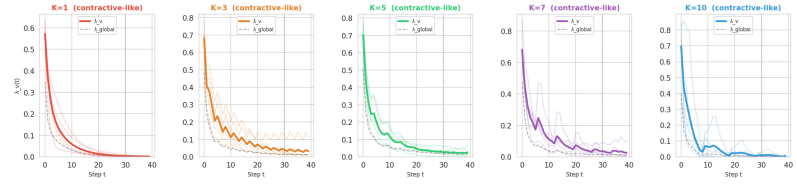


Figure 6: Causal-subspace Lyapunov factor on BabyAI-Unlock. $\lambda_{\mathbf{v}}(t)$ collapses within the first few steps at every K, consistent with no stable latent causal direction.

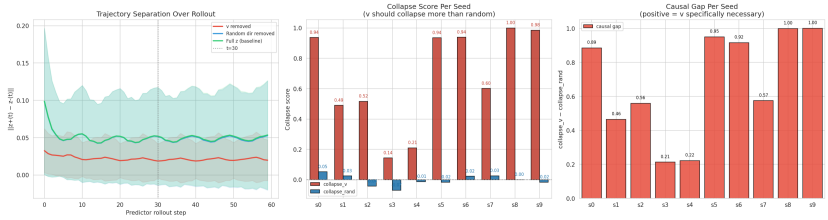


Figure 7: Interventional causal separation on DelayComparison at $K{=}5$. Removing the single estimated direction produces partial collapse; Table 3 shows two to four directions recover it.

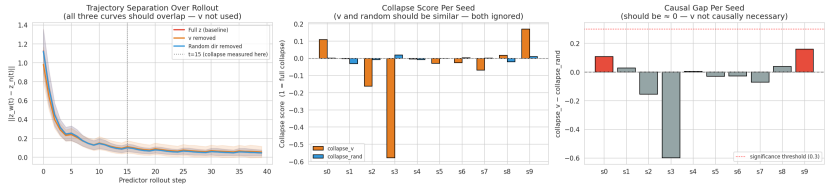


Figure 8: Interventional causal separation on BabyAI at $K{=}5$. Removing $\mathbf{v}$ has no effect, consistent with no concentrated causal direction.