LOW-RANK DYNAMICS-EFFECTIVE LATENT CARRIERS FOR COUNTERFACTUAL ROLLOUT IN LEARNED WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

We ask whether a compact, directly addressable hidden-state intervention can place a learned world model on an intended counterfactual future and then let the model's own dynamics carry that future forward. In a controlled two-object collision environment, we study a 192-dimensional recurrent world model and construct candidate intervention carriers from training-only counterfactual-minus-factual hidden differences. An affine map predicts carrier coordinates from the factual state and requested edit without access to the native counterfactual hidden state at test time. For bounded Single velocity edits, rank 4 is the smallest tested rank satisfying the registered development criteria. A one-shot rank-4 patch launches a 12-transition autonomous rollout, and the frozen procedure satisfies a preregistered 2-of-3 fresh-checkpoint replication rule and remains reusable at nearby anchors. The same Single-derived carrier and Single-only map also support bounded same-object Joint requests. Across matched training regimes, broader counterfactual intervention support is associated with better Joint rollout fidelity and more additive hidden responses. Carrier-relative diagnostics show that the rank-4 subspace is not dynamically closed: its effect couples strongly to the wider recurrent state. These results support a compact intervention-entry interface rather than an intrinsic four-dimensional state, and suggest addressable latent interfaces as a potential design target for controllable world models.

1 INTRODUCTION

World models have evolved from compact learned simulators and latent planning to imagination-based control and action-conditioned generative models (Ha & Schmidhuber, 2018; Hafner et al., 2019; 2023; Bruce et al., 2024; Assran et al., 2025). This advance makes latent state itself a scientific object: not only whether a model predicts, but what its internal state retains and whether that information can guide future behavior. Functional views of representation learning make the same distinction: a useful latent state should preserve what is needed for prediction, control, memory, planning, or intervention, rather than merely expose human-interpretable coordinates (Gelada et al., 2019; K.W. Kim, 2026; Thorpe et al., 2026).

Physical dynamics sharpen the problem. Benchmarks such as Physion establish that learned models can make useful physical predictions (Bear et al., 2021), while recent interpretability studies show that physical information can be distributed, decodable, and causally steerable in world and video models (Geiger et al., 2024; Canby et al., 2025; Zhang, 2026; Alam, 2026). Yet readout accuracy or a short-lived steering effect does not show that an edited hidden state can sustain the intended future under the model's own transition dynamics. We therefore ask a stricter question: can a compact hidden-state intervention place a recurrent world model on an intended counterfactual trajectory and then let the model autonomously continue it?

We study this question in a controlled two-object collision environment with a 192-dimensional recurrent world model. Bounded edits to one velocity component are termed Single, and same-object edits to both velocity components are termed Joint. After first verifying that the model can natively propagate an edited anchor state, we construct checkpoint-specific candidate carriers from training-only counterfactual-minus-factual hidden differences and learn an affine addressable map

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.

from the factual physical state and requested edit to carrier coefficients. At test time, the predicted low-rank patch is applied once, followed by 12 autonomous transitions with no future observations, teacher forcing, repeated hidden correction, or physical-state clamping. Success additionally requires registered target, preservation, physical-validity, and specificity controls.

For Single edits, rank 4 is the smallest tested rank on the preregistered grid that satisfies the development criteria. After rank and evaluation rules are frozen, the procedure satisfies the preregistered 2-of-3 fresh-checkpoint replication rule, remains reusable at nearby anchors, and passes the registered specificity and integrity controls. The same Single-derived carrier and Single-only affine map also support bounded same-object Joint requests without using Joint examples to fit the map. These results identify a compact, directly addressable intervention interface rather than an intrinsic four-dimensional state.

Two further findings sharpen both the mechanism and its possible practical relevance. First, under matched training budgets, broader counterfactual intervention support is associated with better Joint rollout fidelity and more additive native Joint hidden responses, suggesting that richer intervention support may help organize more controllable latent dynamics. Second, carrier-relative diagnostics show that composition-related structure is enriched in the rank-4 subspace but is not confined to it, and local recurrent operators show strong one-step coupling from the carrier into the remaining hidden state. The rank-4 carrier is therefore better viewed as a compact entry interface into high-dimensional recurrent dynamics than as a closed reduced state.

This distinction gives the low-rank interface a potential role beyond interpretation. A compact addressable interface provides a low-dimensional handle for functionally auditing learned state organization and, if analogous structures persist in richer models, could support efficient counterfactual branching or planning without directly manipulating the full hidden state. The present study does not demonstrate such downstream gains; rather, it provides a controlled existence test and suggests addressability and dynamical effectiveness as possible design targets for future world models.

Our contributions are threefold. First, we introduce a stringent dynamics-effective intervention assay in which a one-shot hidden edit must launch a target-specific autonomous counterfactual rollout rather than merely change an immediate readout. Second, we identify and replicate a rank-4 addressable carrier for bounded Single velocity edits and show local temporal reuse and bounded Joint generalization with the same Single-derived interface. Third, matched training-regime and mechanistic analyses connect richer intervention support to improved Joint behavior while showing that the compact carrier is an entry interface, not a closed four-dimensional dynamical state.

2 Related Work

Our work lies at the intersection of functional latent-state sufficiency, causal intervention in distributed representations, and physical world-model analysis. The common question is not only what a hidden state encodes, but whether a compact part of it can be manipulated to produce a specific downstream behavior.

2.1 Latent state, sufficiency, and action relevance

Latent-state work emphasizes preserving task-relevant information rather than human-interpretable coordinates. DeepMDP preserves transition and reward structure for control (Gelada et al., 2019), action-sufficient representations extend this view to partial observability (Huang et al., 2022), and recent formulations define sufficiency relative to prediction, planning, memory, or intervention requirements (K.W. Kim, 2026; Thorpe et al., 2026). Training signals also shape action-relevant state: temporal prediction and explicit physical grounding have been associated with more useful latent structure (Yeom et al., 2026; Yan et al., 2026). Our primary question is complementary: given an already trained recurrent world model, does it expose a compact, directly addressable intervention interface? Matched training-regime comparisons are used only as descriptive evidence about how intervention-support diversity is associated with that structure.

2.2 From probing to causal intervention in neural representations

Probing identifies decodable information, but decodability does not establish causal use (B. Kim et al., 2018; Hewitt & Liang, 2019). Intervention methods therefore remove or alter internal representations, including INLP, Amnesic Probing, LEACE, Concept Bottleneck Models, activation engineering, and interchange interventions (Ravfogel et al., 2020; Elazar et al., 2021; Belrose et al., 2023; Koh et al., 2020; Turner et al., 2023; Geiger et al., 2022). Distributed Alignment Search further allows causal variables to occupy distributed subspaces (Geiger et al., 2024), while causal-probing work stresses selectivity (Canby et al., 2025). We add a recurrent criterion: a one-shot edit must launch the intended autonomous future and remain distinguishable from no-op, random, and wrong-target controls. The Joint extension further asks whether an interface fitted from Single edits can support a bounded two-component request without Joint examples in the addressable map.

2.3 Physical reasoning and object-centered dynamics

Physical reasoning provides a stringent setting for this test. Structured models explicitly encode objects and relations (Battaglia et al., 2016; Wu et al., 2022; Villar-Corrales, Wahdan, & Behnke, 2023; Song et al., 2025), while benchmarks such as IntPhys, CLEVRER, Physion, and IntPhys 2 evaluate physical behavior without identifying the internal mechanism (Riochet et al., 2018; Yi et al., 2019; Bear et al., 2021; Bordes et al., 2025); predictive video models can also acquire intuitive-physics behavior without an explicit physics engine (Garrido et al., 2025). More directly, recent studies probe distributed physical representations or steer predictions through learned directions (Joseph et al., 2026; Zhang, 2026; Alam, 2026), and CRONOS and GAUGE extend evaluation toward counterfactual consistency and measurement-grounded physical fidelity (Begiristain, Dünkel, & Kortylewski, 2026; Wang et al., 2026). Our study targets a different level: a compact hidden-state edit is addressed from the factual state and requested intervention, applied once, and then judged by whether the recurrent model autonomously follows the intended counterfactual future. We also test whether the resulting low-rank carrier is a closed dynamical state or a compact entry interface into the wider hidden dynamics.

3 Method

We study whether a compact hidden-state intervention can place a recurrent world model on an intended counterfactual future and let the model's own dynamics carry that future forward. The controlled environment contains two circular objects moving in a deterministic two-dimensional arena; the model receives noisy numeric object-state observations and maintains a 192-dimensional recurrent hidden state. The primary confirmatory family, MF1, is trained on factual trajectories plus Single counterfactual support. Matched extension families MF0 (factual-only) and MF2 (factual + Single + Joint counterfactual support) use the same architecture and matched optimization budget. Full simulator, model, dataset, and training details are in Appendices A–C.

3.1 Counterfactual setting and intervention objective

A **Single** request changes one velocity component of one object at anchor time t while leaving all other primitive anchor coordinates fixed. A **Joint** request changes both velocity components of the same object. Factual and counterfactual branches share the same pre-anchor history and realized observation noise. For a Single edit,

$$\mathbf{s}_t^{CF} = \mathbf{s}_t^{F} + \delta\, \mathbf{e}_{j,a} \tag{1}$$

where δ is the signed velocity change and $\mathbf{e}_{j,a}$ selects the edited object and velocity axis. Joint requests use the same construction with two same-object velocity components. Simulator rollout from the edited anchor state provides the counterfactual reference trajectory.

We firstly verify that the model can natively represent the requested counterfactual. G1 assimilates the edited anchor observation through the model's ordinary update and then releases the resulting native counterfactual hidden state; G2 adds the complete native counterfactual-minus-factual hidden

difference to the factual hidden state and checks the injection interface. G1 and G2 are prerequisites: failure of either would prevent interpreting a later low-rank result (Appendices C, F, and G).

The intervention itself is applied once to the factual anchor hidden state. It is called dynamics-effective only if the released model then follows the intended counterfactual for 12 autonomous transitions while satisfying registered direct-target, downstream-rollout, preservation, physical-validity, and specificity requirements. After the patch there are no future observation reads, teacher forcing, repeated hidden-state corrections, physical/hidden clamping, or future oracle states; all evaluated actions are zero. Exact metrics and thresholds are in Appendix F.

3.2 Low-rank carrier and addressable patch

Carrier fitting is checkpoint-specific and uses only Single carrier-fit units. For unit i, let $\Delta\mathbf{z}_i = \mathbf{z}_i^{CF} - \mathbf{z}_i^{F}$ be the native counterfactual-minus-factual hidden difference at the anchor. We stack these raw differences and use an uncentered SVD,

$$\mathbf{D} = \left[\Delta\mathbf{z}_1^\top; \ldots; \Delta\mathbf{z}_N^\top\right] = \mathbf{L}\mathbf{\Sigma}\mathbf{V}^\top \qquad \mathbf{U}_r = \mathbf{V}_{[:,1:r]} \tag{2}$$

The columns of $\mathbf{U}_r$ span the candidate rank-r carrier; they are not assigned physical meanings. Projecting a unit's native hidden difference onto $\mathbf{U}_r$ gives a privileged oracle-capacity reference, because it uses that same unit's native counterfactual hidden state. The primary test instead requires addressability without that information.

We therefore fit an affine ridge map $f_{addr,r}$ on Single carrier-fit examples only. Its inputs are the factual primitive state and the requested velocity-edit vector, and its output is the carrier coefficient vector. At evaluation time,

$$\widehat{\mathbf{c}} = f_{addr,r}\left(\mathbf{s}_t^F, \mathbf{e}\right) \qquad \mathbf{z}_t^* = \mathbf{z}_t^F + \beta\mathbf{U}_r\widehat{\mathbf{c}} \tag{3}$$

The patch is applied once; the map is not consulted again during rollout. The primary Single and later Joint addressable routes use $\beta = 1$. For Joint evaluation, the same rank-4 carrier and the same Single-only affine map are used with two nonzero same-object edit components. No Joint data are used to construct the carrier or fit the affine mapper. The actual Joint rollout uses the full affine prediction. Appendix D gives the carrier construction and oracle projection, and Appendix E gives the ridge fit and affine Joint identity.

3.3 Evaluation, controls, rank selection, and frozen replication

The primary Single assay compares the intended patch with factual and simulator references, native-counterfactual and full-hidden references, sham/rank-zero, independent random equal-norm perturbations, wrong-object and wrong-time edits, and leakage plus identity/order checks. These controls separate a target-specific dynamics-effective intervention from generic hidden sensitivity. The Joint extension uses matched no-patch, random equal-norm, wrong-object, and wrong-vector controls (Appendices G and H).

Rank is selected only on three independently trained development checkpoints. The registered candidate grid is $\mathcal{R} = \{1, 2, 4, 8, 12, 16, 20, 32\}$; rank 64 is diagnostic, rank 0 is a no-patch reference, and rank 192 is the full-hidden reference. Single evaluation is separated into S1, where neither factual nor counterfactual trajectories contain a detected future contact step, and S2, where the factual trajectory contains a future contact and the edit produces a sufficiently large post-anchor factual–counterfactual trajectory difference. The strata are never pooled. The selected rank is

$$r^* = \min\left\{ r \in \mathcal{R} : \sum_{q=1}^{3} \mathbb{I}\left(Pass_{r,q}^{S1} \wedge Pass_{r,q}^{S2}\right) \geq 2 \right\} \tag{4}$$

Thus r^* is the smallest tested rank satisfying both strata in at least two of three development checkpoints; it is not an intrinsic-dimension estimate. Once selected, the rank, intervention strength, feature definition, standardization, ridge setting, thresholds, and panel rule are frozen.

Replication then uses three fresh checkpoints trained independently from scratch with the same architecture, dataset, and frozen training procedure but new random seeds; none participates in method development or rank selection. Each fresh checkpoint fits its own carrier and addressable map from its own carrier-fit split, while its confirmatory split is evaluation-only. No new rank sweep, mapper comparison, threshold change, checkpoint exclusion, or post-hoc retuning is allowed. Checkpoints, rather than individual trajectory units, are the replication level.

3.4 Temporal, Joint, and mechanistic extensions

After establishing the Single rank-4 interface at $t = 7$, two temporal tests examine whether the interface works at nearby times. B1 refits the frozen procedure independently at anchors $t = 5, 6, 7, 8, 9$ and evaluates each interface at its own anchor. B2 transports the exact $t = 7$ basis, affine-map parameters, feature statistics, rank, and intervention strength to $t = 5, 6, 8, 9$ without refitting or restandardization. These tests assess local recoverability and direct reuse, not globally time-invariant coordinates (see Appendix G for more details).

Joint analyses are performed at $t = 7$ on identity-matched factual/Single-x/Single-y/Joint quadruplets. The registry contains 512 base units, split into 411 FIT units for factual-reference whitening and local-operator fitting and 101 held-out TEST units for composition, route, and specificity comparisons. MF1 and MF2 compare Native Joint, a privileged projection of the native Joint hidden difference onto the frozen Single-derived $\mathbf{U}_4$, and the Single-only addressable route on the same TEST units. MF0 contributes to native Joint and composition comparisons but not to Single-derived oracle/addressable routes. The MF0/MF1/MF2 families keep architecture, optimizer, update count, and total branch presentations matched; only intervention-support composition differs, so family comparisons are interpreted descriptively rather than as universal causal effects of training.

For composition, we compare the native Joint hidden response with the sum of the matched Single responses after checkpoint-specific factual-reference whitening; static composition is primary and recurrence-based composition is supporting. Mechanistic diagnostics then express $\mathbf{U}_4$ and its complement in the same whitened coordinates and fit local route-conditioned affine one-step operators from FIT trajectories. Projector blocks quantify carrier retention and leakage into the complement. These operators are descriptive local fits, not exact recurrent dynamics, global Koopman operators, or reduced-state models (more details presented in Appendices H and I).

All positive claims remain bounded to the tested state, edit, time, model, and horizon supports. In particular, rank 4 denotes the smallest passing tested rank for the registered Single procedure; it does not imply four named physical variables, an intrinsic four-dimensional state, or dynamical closure. Successful Joint editing establishes bounded same-object compositional addressability, not arbitrary multi-component or multi-object control.

4 Results

We organize the results around four questions: whether a compact addressable carrier exists and replicates, whether it can be reused across time and extended to Joint requests, how matched intervention support is associated with Joint behavior, and whether the resulting rank-4 interface is dynamically closed.

4.1 Rank 4 is the smallest tested addressable carrier and replicates

Before any low-rank claim, the factual, native-counterfactual, and full-hidden-equivalence reference routes passed the required S1/S2 adequacy checks on both development and fresh checkpoints; the full-hidden route also reproduced the native counterfactual rollout within numerical tolerance. The model and injection interface can therefore represent the requested counterfactual before low-rank reproduction is tested.

On the preregistered development grid, ranks 1 and 2 failed, while rank 4 was the smallest tested rank that passed both strata in at least two of three checkpoints together with the mandatory controls, as shown in Figure 1. A privileged oracle-projection sweep reached the same smallest passing tested rank, as shown in Figure D1. But the hidden-difference spectrum is only descriptive. Thus rank 4 is a behavioral result of the registered assay, not an intrinsic-state dimension estimate.

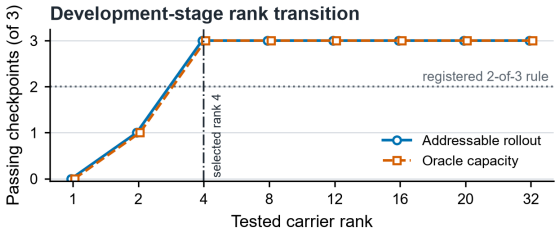


Figure 1: Development-stage rank transition for the primary Single intervention. Rank 4 is the smallest tested rank satisfying the registered panel rule and controls for both the addressable rollout and the privileged oracle-capacity route. Rank selection is based on rollout behavior and controls, not on the descriptive carrier spectrum reported in the Appendix D.

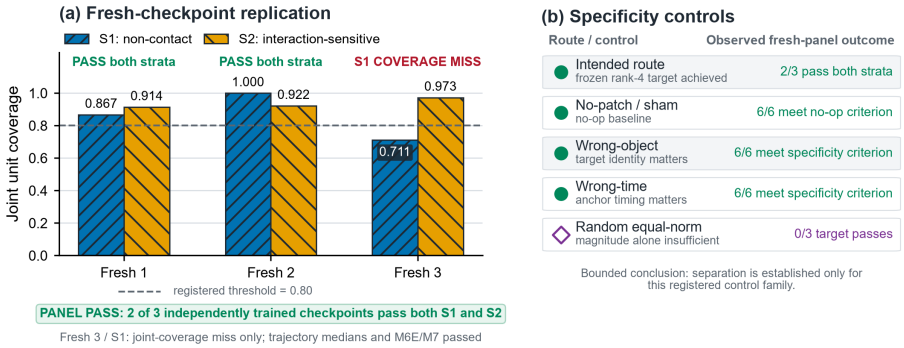


Figure 2: Fresh-checkpoint replication and specificity controls for the frozen rank-4 addressable carrier. (a) Unit-level joint-pass coverage in S1 and S2; Fresh 1 and Fresh 2 pass both strata, whereas Fresh 3 misses only S1, satisfying the registered 2-of-3 rule. (b) Registered reference and specificity-control outcomes; a passing control means that it behaves according to its own registered criterion.

After freezing the rank, intervention strength, map family, standardization, thresholds, and panel rule, the procedure satisfied the preregistered 2-of-3 fresh-checkpoint rule as shown in Figure 2. Unit-level joint-pass coverage in S1/S2 was 0.867/0.914, 1.000/0.922, and 0.711/0.973 for Fresh 1–3, respectively: Fresh 1 and 2 passed both strata, whereas Fresh 3 missed only S1. Sham, wrong-object, wrong-time, leakage, and identity/order controls behaved as registered, and independent random equal-norm perturbations produced 0/3 fresh-model target passes. The frozen procedure therefore replicates with checkpoint-level heterogeneity rather than uniform success.

4.2 The interface is reusable across nearby anchors and supports bounded Joint requests

The established Single interface is locally reusable in time. As shown in Figure G1, refitting the frozen rank-4 procedure separately at $t = 5 - 9$ yielded the registered 2-of-3 outcome at every anchor; more strongly, transporting the exact t=7 basis, affine-map parameters, feature statistics, rank, and intervention strength unchanged to $t = 5, 6, 8, 9$ also passed the 2-of-3 rule without target-anchor refitting or restandardization. These results support reuse over the tested nearby window, not globally time-invariant latent coordinates.

The same Single-derived rank-4 carrier and Single-only affine map also support bounded same-object Joint requests that edit both velocity components. On the same 101 held-out units per checkpoint, family-median joint-pass coverage for Native/oracle/Addressable routes was 0.970/0.941/0.901 in MF1 and 0.990/1.000/0.980 in MF2; addressable rollout RMSE was 0.157 and 0.125, respectively, as shown in Figure 3. Matched no-patch, random equal-norm, wrong-object, and wrong-vector controls did not reproduce the intended effect. Because the map is fitted only from Single examples, this establishes bounded compositional addressability, not an arbitrary additive control algebra.

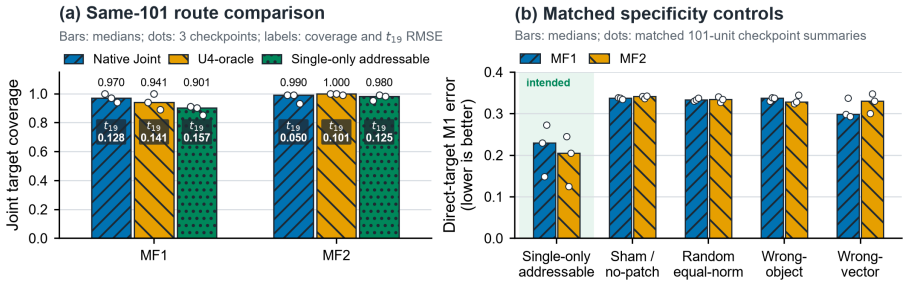


Figure 3: Bounded same-object Joint addressability from a Single-derived rank-4 interface. (a) Native Joint, U_4-oracle, and Single-only addressable routes on the same held-out units in MF1 and MF2. Bars show family medians and dots show checkpoint-level values. (b) Matched specificity controls for the addressable Joint route; lower direct-target error is better.

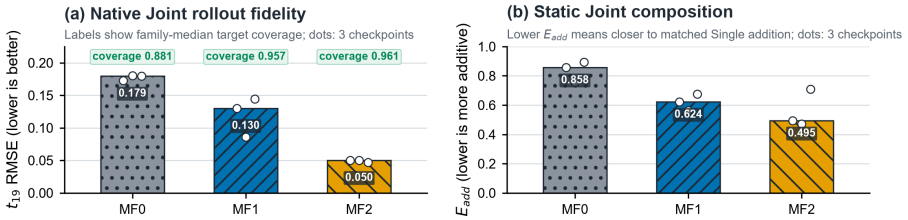


Figure 4: Native Joint rollout fidelity and static hidden-response composition across matched training families. (a) Native Joint terminal RMSE, with family-median joint-pass coverage reported above the bars. (b) Static composition error $\mathbf{E_{add}}$; lower values indicate closer agreement between the native Joint response and the additive combination of matched Single responses. Dots show checkpoint-level values.

4.3 Broader intervention support is associated with better Joint fidelity and composition

To ask whether training support is associated with this behavior, we compare MF0 (factual only), MF1 (factual + Single-counterfactual), and MF2 (factual + Single- and Joint-counterfactual) under matched architecture and training budgets. Native Joint competence is already substantial in MF0, with family-median joint-pass coverage 0.881/0.957/0.961 across MF0/MF1/MF2, but rollout fidelity improves more strongly: terminal RMSE decreases from 0.179 to 0.130 to 0.050 (as shown in Figure 4a).

Native Joint responses also become more additive: static composition error decreases from 0.858 to 0.624 to 0.495 (as shown in Figure 4b). Dynamic composition shows the same ordering weaklier, and checkpoint ranges overlap, so we interpret these as family-level associations rather than a universal causal law. Within this bounded setting, the result suggests that broader counterfactual intervention support may improve not only rollout fidelity but also the organization of compositional hidden responses.

4.4 The rank-4 carrier is a compact entry, not a closed state

Carrier-relative diagnostics show that the rank-4 interface is not a closed four-dimensional state. The shift toward greater additivity occurs both inside the carrier and in its orthogonal complement, yet composition-defect squared norm is strongly enriched in the four-dimensional carrier: 30.4% in MF1 and 17.5% in MF2, corresponding to about 14.6-fold and 8.4-fold enrichment over the dimensional baseline (as shown in Figure 5a).

Local recurrent fits give the stronger boundary. Family-median one-step carrier retention is only 0.016 for Single and Joint in MF1 and 0.027/0.028 in MF2, with complementary leakage of 0.972–0.984 (as shown in Figure 5b). Thus the compact carrier provides an effective intervention entry, but its one-step effect is dominated by coupling into the wider recurrent hidden state. Supporting geometry

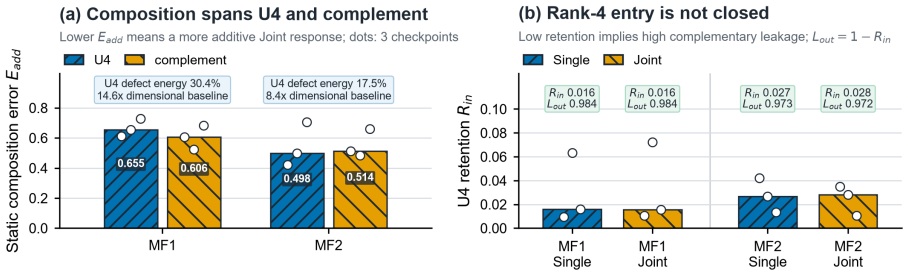


Figure 5: Carrier-relative composition and recurrent unfolding of the Single-derived rank-4 interface. (a) Static composition error inside the whitened $\mathbf{U}_4$ carrier and its orthogonal complement; annotations report composition-defect fraction and enrichment relative to the 4/192-dimensional baseline. (b) One-step carrier retention R_{in} for the fitted local Single and Joint operators; low retention implies high complementary leakage.

diagnostics likewise show that better MF2 Joint behavior does not correspond to smaller distance from the projected native-Joint oracle.

A separate position-edit stress test provides a negative contrast: raw rollout success was reproduced by no-patch and random controls, while wrong-object specificity failed, as shown in Figure G2. This reinforces the central evidence standard: successful steering or geometric proximity is insufficient unless a one-shot intervention produces the targeted autonomous future and separates from plausible controls.

5 Discussion

The results support a compact, checkpoint-specific intervention interface for the tested velocity counterfactuals: a low-rank patch can be addressed from the factual state and requested edit, applied once, and then propagated by the model's own recurrent dynamics. The Joint and mechanistic analyses sharpen what this means—and what it does not.

5.1 What the rank-4 interface means

Dynamics-effective is a functional notion. Decodability, immediate steering, subspace capacity, or geometric proximity do not by themselves show that an edited representation supports the intended future. Our criterion instead requires a one-shot, test-time-addressable patch to launch the target counterfactual, survive autonomous rollout, and separate from no-op, random, wrong-target, and integrity controls. The bounded Joint result extends this criterion beyond isolated one-component edits: the same Single-derived rank-4 carrier and Single-only map support same-object two-component requests without fitting Joint examples in the map. This does not imply an arbitrary additive control algebra or multi-object compositionality.

The carrier-relative diagnostics also rule out a stronger interpretation. Composition-related structure is enriched in the rank-4 carrier but is not confined to it, and only a few percent of carrier-originating one-step operator strength remains inside the carrier. Rank 4 should therefore be read as the smallest passing tested intervention interface for the registered Single assay, not as an intrinsic state dimension or a closed four-dimensional dynamical state.

5.2 Training and practical implications

The matched MF0/MF1/MF2 comparison provides a separate training-design clue. Under matched architecture and training budgets, broader counterfactual intervention support is associated with higher Joint rollout fidelity and more additive Joint hidden responses. Because checkpoint ranges overlap and the comparison spans only these families, we treat this as a family-level association rather than a universal causal law. Still, it suggests that diversity of intervention support may influence not only behavioral accuracy but also how compositional response structure is organized in recurrent state.

This gives the low-rank interface a potential practical role despite the study's basic-research focus. A compact addressable interface could provide a lower-complexity control surface for branching a learned world model into candidate counterfactual futures, and the dynamics-effective assay itself offers a functional audit of whether latent changes are target-specific and actually used by the transition dynamics. These possibilities are not demonstrated here as planning or control improvements; rather, they motivate future work on training world models whose latent states are deliberately more addressable, compositional, or dynamically well organized.

5.3 Scope and next steps

The claims remain local to the tested setting. Carrier bases are checkpoint-specific; temporal reuse is established only over the nearby tested anchors; Joint requests are bounded and same-object; and the environment uses one recurrent architecture, two objects, deterministic dynamics, numeric observations, zero evaluated actions, and a 12-transition horizon. The results therefore do not establish the same structure for raw-video models, stochastic or partially observed dynamics, active control, larger scenes, or real systems. The position-edit negative contrast further shows that raw rollout improvement is not sufficient when no-patch or random perturbations behave similarly.

A fuller effective state would require more than compact intervention entry: it would also need predictive sufficiency and some form of dynamical closure, and under partial observability may require history- or belief-like information. Testing these properties across architectures and scales, and asking whether they can be encouraged during training, are natural next steps.

6 Conclusion

We identify a compact, directly addressable intervention interface in a 192-dimensional recurrent world model. For bounded Single velocity edits, rank 4 is the smallest tested rank satisfying the registered development criteria; a one-shot patch launches a 12-transition autonomous rollout, satisfies the preregistered fresh-checkpoint replication rule, and can be reused at nearby anchors. The same Single-derived carrier and Single-only affine map also support bounded same-object Joint requests. Across matched training regimes, broader counterfactual intervention support is associated with better Joint rollout fidelity and more additive hidden responses. Mechanistic diagnostics show, however, that the rank-4 carrier is not dynamically closed: its effect couples strongly to the wider recurrent state. We therefore interpret rank 4 as a compact intervention-entry interface rather than an intrinsic four-dimensional state. These results motivate addressable low-dimensional interfaces as both a functional diagnostic of learned world models and a potential basis for compact counterfactual branching and future controllable world-model design.

AI Assistance Disclosure

Generative AI tools, including ChatGPT and OpenAI Codex, were used as research-assistance tools for literature discovery and synthesis, refinement of research questions and experimental protocols, code implementation and debugging, analysis organization, and manuscript preparation. All experimental protocols, scientific decisions, formal runs, interpretation of results, and final claims were reviewed and approved by the authors. AI-generated code, references, analyses, and text were reviewed by the authors and checked against the relevant source code, experimental records, and primary literature where applicable. The authors take full responsibility for the accuracy and integrity of the work.

REFERENCES

Alam, Nahid. 2026. "Causal Physics Steering in Video World Models via Concept Activation Vectors." *arXiv preprint arXiv:2605.24322*.

Assran, Mido, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, and Artem Zholus. 2025. "V-jepa 2: Self-supervised video models enable understanding, prediction and planning." *arXiv preprint arXiv:2506.09985*.

Battaglia, Peter, Razvan Pascanu, Matthew Lai, and Danilo Jimenez Rezende. 2016. "Interaction networks for learning about objects, relations and physics." *Advances in neural information processing systems* 29.

Bear, Daniel M, Elias Wang, Damian Mrowca, Felix J Binder, Hsiao-Yu Fish Tung, RT Pramod, Cameron Holdaway, Sirui Tao, Kevin Smith, and Fan-Yun Sun. 2021. "Physion: Evaluating physical prediction from vision in humans and machines." *arXiv preprint arXiv:2106.08261*.

Begiristain, León, Olaf Dünkel, and Adam Kortylewski. 2026. "CRONOS: Benchmarking Counterfactual Physical Consistency in Video Models." *arXiv preprint arXiv:2605.23699*.

Belrose, Nora, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. 2023. "Leace: Perfect linear concept erasure in closed form." *Advances in Neural Information Processing Systems* 36: 66044–66063.

Bordes, Florian, Quentin Garrido, Justine T Kao, Adina Williams, Michael Rabbat, and Emmanuel Dupoux. 2025. "Intphys 2: Benchmarking intuitive physics understanding in complex synthetic environments." *arXiv preprint arXiv:2506.09849*.

Bruce, Jake, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, and Chris Apps. 2024. "Genie: Generative interactive environments." Forty-first international conference on machine learning.

Canby, Marc E, Adam Davies, Chirag Rastogi, and Julia Hockenmaier. 2025. "How reliable are causal probing interventions?" Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics.

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

Garrido, Quentin, Nicolas Ballas, Mahmoud Assran, Adrien Bardes, Laurent Najman, Michael Rabbat, Emmanuel Dupoux, and Yann LeCun. 2025. "Intuitive physics understanding emerges from self-supervised pretraining on natural videos." *arXiv preprint arXiv:2502.11831*.

Geiger, Atticus, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah Goodman, and Christopher Potts. 2022. "Inducing causal structure for interpretable neural networks." International Conference on Machine Learning.

Geiger, Atticus, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. 2024. "Finding alignments between interpretable causal variables and distributed neural representations." Causal Learning and Reasoning.

Gelada, Carles, Saurabh Kumar, Jacob Buckman, Ofir Nachum, and Marc G Bellemare. 2019. "Deepmdp: Learning continuous latent space models for representation learning." International conference on machine learning.

Ha, David, and Jürgen Schmidhuber. 2018. "World models." *arXiv preprint arXiv:1803.10122* 2 (3): 440.

Hafner, Danijar, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. 2019. "Learning latent dynamics for planning from pixels." International conference on machine learning.

Hafner, Danijar, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. 2023. "Mastering diverse domains through world models." *arXiv preprint arXiv:2301.04104*.

Hewitt, John, and Percy Liang. 2019. "Designing and interpreting probes with control tasks." Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp).

Huang, Biwei, Chaochao Lu, Liu Leqi, José Miguel Hernández-Lobato, Clark Glymour, Bernhard Schölkopf, and Kun Zhang. 2022. "Action-sufficient state representation learning for control with structural constraints." International Conference on Machine Learning.

Joseph, Sonia, Quentin Garrido, Randall Balestriero, Matthew Kowal, Thomas Fel, Shahab Bakhtiari, Blake Richards, and Mike Rabbat. 2026. "Interpreting physics in video world models." *arXiv preprint arXiv:2602.07050*.

Kim, Been, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, and Fernanda Viegas. 2018. "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)." International conference on machine learning.

Kim, Keon Woo. 2026. "Latent State Design for World Models under Sufficiency Constraints." *arXiv preprint arXiv:2605.01694*.

Koh, Pang Wei, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. 2020. "Concept bottleneck models." International conference on machine learning.

Ravfogel, Shauli, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. "Null it out: Guarding protected attributes by iterative nullspace projection." Proceedings of the 58th annual meeting of the association for computational linguistics.

Riochet, Ronan, Mario Ynocente Castro, Mathieu Bernard, Adam Lerer, Rob Fergus, Véronique Izard, and Emmanuel Dupoux. 2018. "Intphys: A framework and benchmark for visual intuitive physics reasoning." *arXiv preprint arXiv:1803.07616*.

Song, Yeon-Ji, Jaein Kim, Suhyung Choi, Jin-Hwa Kim, and Byoung-Tak Zhang. 2025. "OCK: Unsupervised Dynamic Video Prediction with Object-Centric Kinematics." 2025 IEEE/CVF International Conference on Computer Vision (ICCV).

Thorpe, Adam J, Stepan Tretiakov, Cheng-Hsi Hsiao, Su Ann Low, Xingjian Li, Hassan Iqbal, Neel P Bhatt, Ufuk Topcu, and Krishna Kumar. 2026. "Physically Viable World Models: A Case for Query-Conditioned Embodied AI." *arXiv preprint arXiv:2605.30542*.

Turner, Alexander Matt, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. "Steering language models with activation engineering." *arXiv preprint arXiv:2308.10248*.

Villar-Corrales, Angel, Ismail Wahdan, and Sven Behnke. 2023. "Object-centric video prediction via decoupling of object dynamics and interactions." 2023 IEEE International conference on image processing (ICIP).

Wang, Shuai, Yaxin Feng, Xuekun Jiang, Shihan Tian, Ningyu Yan, Xing Shen, Chaoyang Lyu, Hui Wang, Yunsong Zhou, and Hanqing Wang. 2026. "GAUGE: A Measurement-Grounded Benchmark for Physical Fidelity in Simulation Engines and Video World Models." *arXiv preprint arXiv:2608.05948*.

Wu, Ziyi, Nikita Dvornik, Klaus Greff, Thomas Kipf, and Animesh Garg. 2022. "Slotformer: Unsupervised visual dynamics simulation with object-centric models." *arXiv preprint arXiv:2210.05861*.

Yan, Haodong, Jiaguan Zhu, Mingyuan Jia, Ruiqing Yin, Junjie He, Zhide Zhong, Junfeng Li, Jinxuan Lu, Hengtao Li, Tianran Zhang, Jiayi Chen, Wenxuan Song, Wen Chen, Yuxiang Gao, and Haoang Li. 2026. Is Forward Prediction Enough? Physical State Grounding for JEPA World Models. *arXiv preprint arXiv: 2608.06799*

Yeom, Jewon, Hanseul Kim, Jeongjae Park, Sungmok Jung, Jaejin Lee, and Taesup Kim. 2026. "What Makes Video World Model Latents Action-Relevant: Prediction over Reconstruction." *arXiv preprint arXiv:2606.07687*.

Yi, Kexin, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. 2019. "Clevrer: Collision events for video representation and reasoning." *arXiv preprint arXiv:1910.01442*.

Zhang, Xinyu. 2026. "What Do World Models Learn in RL? Probing Latent Representations in Learned Environment Simulators." *arXiv preprint arXiv:2603.21546*.

A Formal Setup and Notation

This appendix collects the minimal notation needed to interpret Sections 3–4. Environment, model, data, and training details are in Appendix B; exact counterfactual construction and strata in Appendix C; carrier and addressability methods in Appendices D–E; evaluation rules in Appendix F; controls and temporal protocols in Appendix G; Joint analyses in Appendix H; and carrier-relative dynamics diagnostics in Appendix I.

A.1 Core notation

We use the following symbols throughout the supplementary material.

- $\mathbf{s}_{j,t} \in \mathbb{R}^4$ is object j's primitive state; $\mathbf{s}_t \in \mathbb{R}^8$ is the two-object primitive state; and $\mathbf{o}_t \in \mathbb{R}^{2\times 6}$ is the numeric observation.
- $\mathbf{z}_t \in \mathbb{R}^{192}$ is the full recurrent hidden state manipulated by the intervention analysis.
- t is the anchor index and H the number of autonomous future transitions; the main assay uses $H = 12$.
- $\mathbf{e}_{j,a} \in \mathbb{R}^8$ selects one velocity coordinate in the primitive state, whereas $\mathbf{e} \in \mathbb{R}^4$ is the addressable velocity-request vector used by the coefficient map. Single requests have one nonzero slot; Joint requests have two same-object nonzero slots.
- $\mathbf{z}_t^F$ and $\mathbf{z}_t^{CF}$ are the factual and native-counterfactual anchor states, and $\Delta\mathbf{z}_t^{native} = \mathbf{z}_t^{CF} - \mathbf{z}_t^F$.
- $\mathbf{U}_r \in \mathbb{R}^{192\times r}$ is the rank-r carrier basis, $\mathbf{c} \in \mathbb{R}^r$ its coefficient vector, and f_{addr} the addressable map from $\left(\mathbf{s}_t^F, \mathbf{e}\right)$ to $\mathbf{c}$.
- R_H is the autonomous rollout operator, and Ω is the bounded state-edit-time domain over which intervention claims are made.

A.2 Same-history Single and Joint requests

Factual and counterfactual branches share the same pre-anchor history and realized observation noise; Appendix C gives the exact construction.

A Single request changes one velocity coordinate at the anchor:

$$\mathbf{s}_t^{CF} = \mathbf{s}_t^F + \delta v\, \mathbf{e}_{j,a}, \qquad a \in \{x, y\} \tag{A1}$$

Here, δv is the signed velocity increment; $j \in \{A, B\}$ is the edited object; $a \in \{x, y\}$ is the edited velocity axis; and $\mathbf{e}_{j,a} \in \mathbb{R}^8$ is zero except at that velocity coordinate. Superscripts F and CF denote the factual and counterfactual branches.

The selector $\mathbf{e}_{j,a} \in \mathbb{R}^8$ is zero except at the selected velocity coordinate. A Joint request changes both velocity components of the same object:

$$\mathbf{s}_t^{CF} = \mathbf{s}_t^F + \delta v_x\, \mathbf{e}_{j,x} + \delta v_y\, \mathbf{e}_{j,y}, \qquad \delta v_x \neq 0,\ \delta v_y \neq 0 \tag{A2}$$

Here, δv_x and δv_y are the two nonzero velocity changes applied to the same object j; $\mathbf{e}_{j,x}$ and $\mathbf{e}_{j,y}$ select its horizontal and vertical velocity slots.

and the other object's primitive state remain unchanged directly at the anchor. Both Single and Joint edits are restricted to their registered support.

A.3 Low-rank and addressable intervention

A rank-r candidate restricts the one-shot hidden perturbation to the column space of $\mathbf{U}_r$:

$$\delta\mathbf{z} = \mathbf{U}_r\mathbf{c} \tag{A3}$$

Here, $\delta \mathbf{z} \in \mathbb{R}^{192}$ is the one-shot hidden perturbation, $\mathbf{U}_r \in \mathbb{R}^{192\times r}$ is the checkpoint-specific rank-r carrier basis, and $\mathbf{c} \in \mathbb{R}^r$ gives the coordinates of the requested edit in that basis.

Addressability requires the coefficients to be predicted without using the evaluation unit's native counterfactual hidden state:

$$\mathbf{c} = f_{addr}\left(\mathbf{s}_t^F, \mathbf{e}\right) \tag{A4}$$

Here, f_{addr} is the addressable coefficient map, $\mathbf{s}_t^F \in \mathbb{R}^8$ is the factual two-object primitive state, and $\mathbf{e} \in \mathbb{R}^4$ is the requested velocity edit. The output $\mathbf{c}$ is predicted without access to $\mathbf{z}_t^{CF}$.

Exact carrier construction, ridge fitting, and test-time patching are defined in Appendices D–E. In the Joint evaluation, the rank-4 carrier and addressable map remain Single-derived and Single-only fitted.

A.4 Dynamics-effective criterion and bounded scope

The evaluation distinguishes direct targets, quantities derived immediately from the edited anchor state, downstream dynamical consequences, and variables that should remain unaffected. A dynamics-effective intervention must launch the registered counterfactual while satisfying the corresponding preservation, validity, and specificity criteria:

$$\widehat{\mathbf{Y}}^{*}_{t:t+H} = R_H\left(\mathbf{z}_t^{*}\right) \approx \mathbf{Y}^{CF}_{t:t+H} \tag{A5}$$

Here, $\mathbf{z}_t^{\star}$ is the patched factual hidden state, R_H is the autonomous H-transition rollout operator, $\widehat{\mathbf{Y}}^{\star}_{t:t+H}$ is the decoded patched rollout, and $\mathbf{Y}^{CF}_{t:t+H}$ is simulator counterfactual truth. The main assay uses $H = 12$.

The relation in Eq. (A5) denotes satisfaction of the registered multi-part decision rule, not coordinate-wise equality or a single Euclidean-distance threshold. Exact metrics and decision rules are in Appendix F, with route and control definitions in Appendices G–H.

Positive intervention claims are restricted to the tested operational domain

$$\Omega = \Omega_{state} \times \Omega_{int} \times \Omega_{time} \tag{A6}$$

Here Ω_{state} is the tested factual-state region, Ω_{int} the registered edit family and magnitude support, and Ω_{time} the evaluated anchor times. The primary Single assay uses $t = 7$, with separate temporal tests over $t \in \{5, 6, 7, 8, 9\}$; the Joint extension is evaluated at $t = 7$ within its matched support. Position edits form a separate diagnostic family.

B Environment, Observation Interface, Model Architecture, and Training

This appendix gives the implementation details needed to reproduce the controlled environment, observation interface, 192-D recurrent carrier, training protocol, checkpoint roles, and matched MF0–MF2 training families used in the experiments.

B.1 Simulator and physical state

Table B1: Simulator and physical-state specification

Item	Implemented value
Scene	Two circular objects moving in a two-dimensional square reflecting arena
Primitive state per object	x, y, horizontal velocity, vertical velocity
Two-object primitive state	2×4 array; 8 scalar values
Time step	0.25
Object radius	0.25

Item	Implemented value
Valid object-center coordinate range	[-4.75, 4.75] on each axis
Process noise	None; simulator is deterministic
Stored episode	22 physical states (indices 0–21) and 21 action slots
Explicit mass / restitution	Not represented as simulator parameters

Within each 0.25-step, the deterministic simulator solves for the earliest circle–circle contact time from the relative-motion quadratic, advances to contact, applies the implemented normal impulse update, and then advances the remaining time; wall crossings are reflected. This is consistent with equal-mass frictionless normal-collision behavior, although mass and restitution are not explicit simulator variables.

B.2 Observation interface

Table B2: Observation interface

Property	Value
Per-time observation shape	2×6
Per-object channels	normalized noisy x, y, horizontal velocity, vertical velocity, visibility, identity
Identity values	-1 for one object slot and +1 for the other
Directly observed when visible	position, velocity, identity
Not directly observed	radius, contact
Position-noise standard deviation	0.005
Velocity-noise standard deviation	0.01
Visibility schedule	frames 0–4 and the selected anchor are visible; physical channels at other frames are zeroed

The visibility schedule keeps frames 0–4 and the selected anchor visible, while physical observation channels after frame 4 and before the selected anchor are zeroed. Thus, for the main anchor $t = 7$, frames 5 and 6 form a two-step unobserved interval; the recurrent state is propagated across this interval before the visible anchor observation is assimilated. Factual/counterfactual pairs share the same realized observation-noise array. When visible, position, velocity, and identity are supplied directly; radius and contact are not observed. Thus carrier differences are not driven by independent noise draws, and the study does not claim visual or finite-difference velocity inference.

B.3 Recurrent world model

The deterministic belief-like recurrent model has 496200 trainable parameters and exposes the complete 192-D carrier listed in Table B3. Figures B1–B3 show input encoding, the deterministic recurrent/prior update, posterior assimilation, and final carrier assembly.

Table B3: Recurrent carrier composition and model-module dimensions

Carrier block	Width
Object 0 deterministic recurrent state	48
Object 0 middle component	16
Object 1 deterministic recurrent state	48
Object 1 middle component	16
Global deterministic recurrent state	64
Total	192
Module	**Implemented dimensions**
Observation encoder	$6 \to 128 \to 64$
Action encoder	$4 \to 128 \to 64$
Message network	$192 \to 256 \to 128 \to 64$
Shared object GRU	input 144, hidden 48

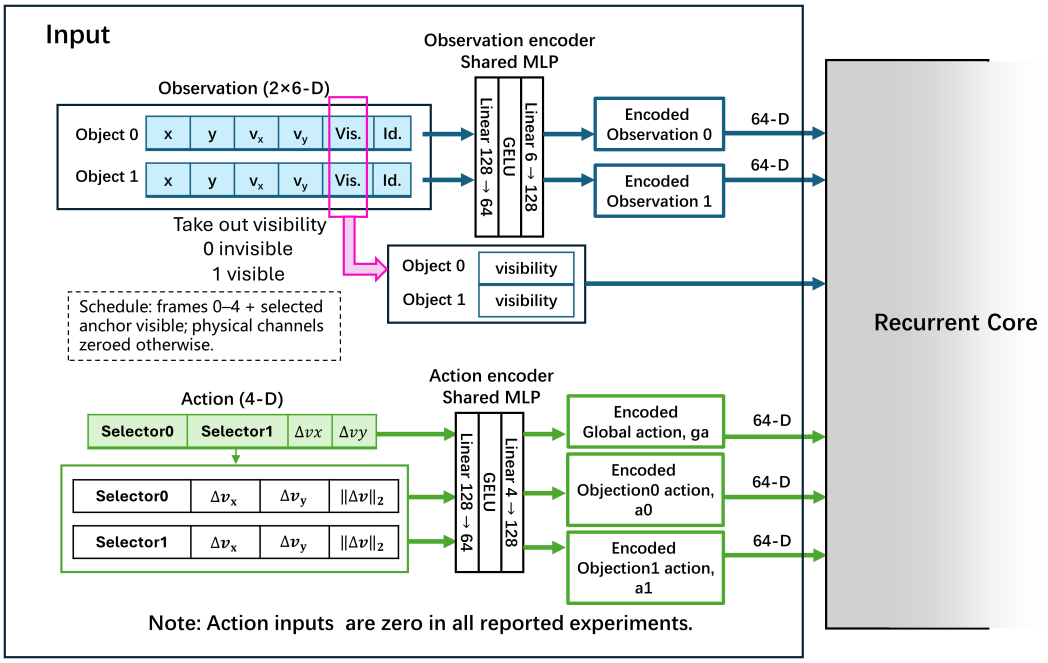


Figure B1: Observation and action encoding. The shared observation encoder produces one 64-D embedding per object and retains visibility separately. Frames 0–4 and the selected anchor are visible; physical observation channels are zeroed at nonvisible frames. The action encoder produces two object-specific and one global 64-D embedding. All reported actions are zero.

Carrier block	Width
Global GRU	input 192, hidden 64
Prior network	$176 \to 256 \to 128 \to 16$
Posterior network	$176 \to 256 \to 128 \to 16$
Decoder	$192 \to 384 \to 192 \to 8$
Normalization layers	none

B.3.1 Observation and action encoding

A shared $6 \to 128 \to 64$ MLP encodes each object observation while visibility is passed separately for prior/posterior selection. The action encoder produces three 64-D embeddings: two object-specific embeddings and one global embedding. All reported experiments use zero actions.

B.3.2 Deterministic recurrent and prior update

The previous-step carrier is unpacked and the shared message MLP is applied to the ordered and object-swapped object states. For object i, the Object GRUCell receives $[\mathbf{m_i},\ \mathbf{message_i},\ \mathbf{a_i}]$ as a 144-D data input and $\mathbf{h_i}$ as a separate 48-D recurrent hidden state, producing $\mathbf{h'_i}$. The Global GRU receives $[\mathbf{h'_0},\ \mathbf{m_0},\ \mathbf{h'_1},\ \mathbf{m_1},\ \mathbf{ga}]$ as its 192-D data input and $\mathbf{gs}$ as its 64-D recurrent hidden state, producing $\mathbf{gs'}$. The shared prior MLP maps $[\mathbf{h'_i},\ \mathbf{gs'},\ \mathbf{a_i}]$ to a capped 16-D prior middle component.

B.3.3 Posterior assimilation and carrier assembly

When an anchor observation is available, the shared posterior MLP maps $[\mathbf{h'_i},\ \mathbf{gs'},\ \mathbf{e_obs_i}]$ to a 16-D posterior middle component. With visibility $vis_i \in \{0, 1\}$, the selected middle component is

$$\mathbf{m'_i} = vis_i\, \mathbf{m_i^{post}} + (1 - vis_i)\, \mathbf{m_i^{prior}}$$

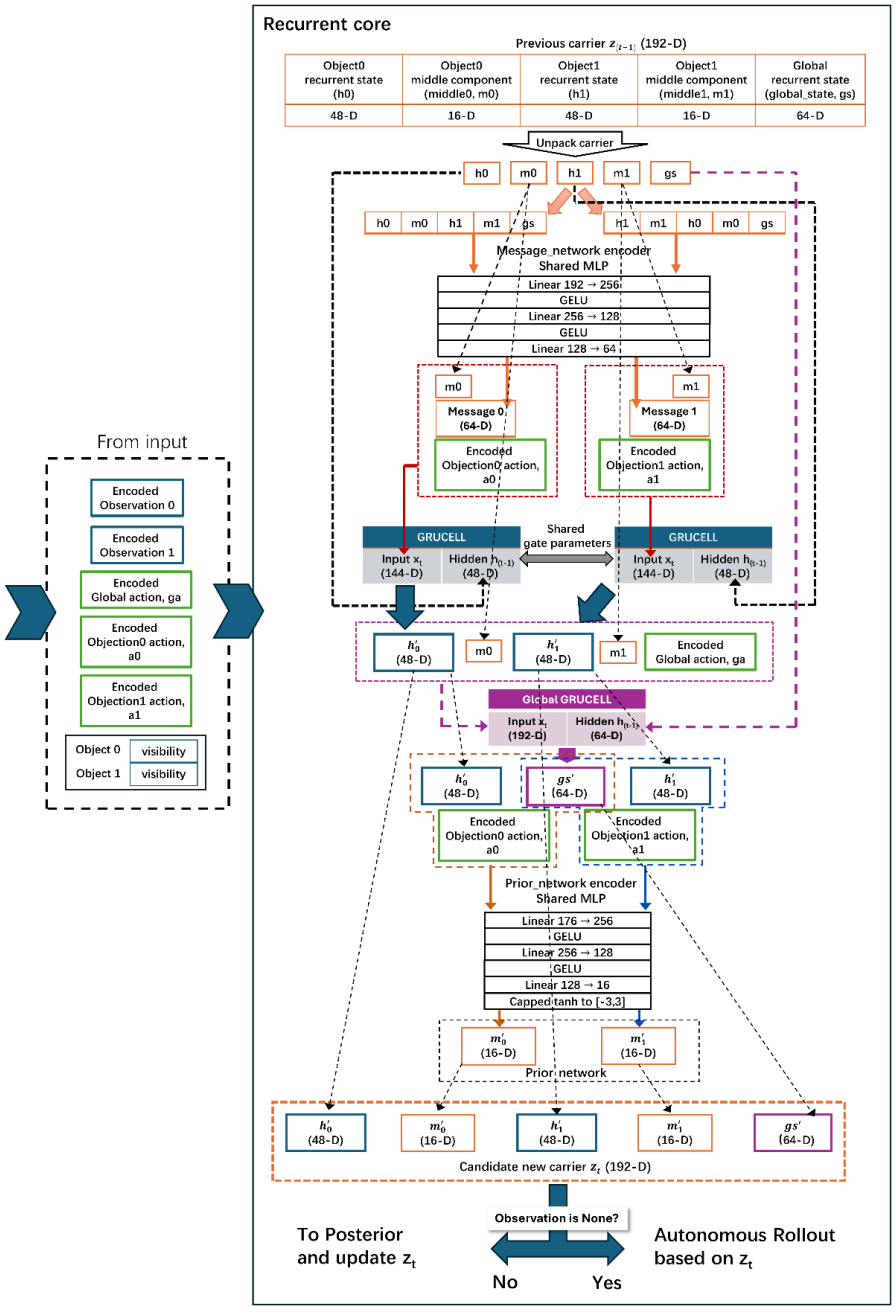


Figure B2: Deterministic recurrent and prior update. The message MLP and Object GRU Cells are shared across object slots; the Global GRU produces $\mathbf{gs}'$, and the shared prior MLP produces the two 16-D prior middle components.

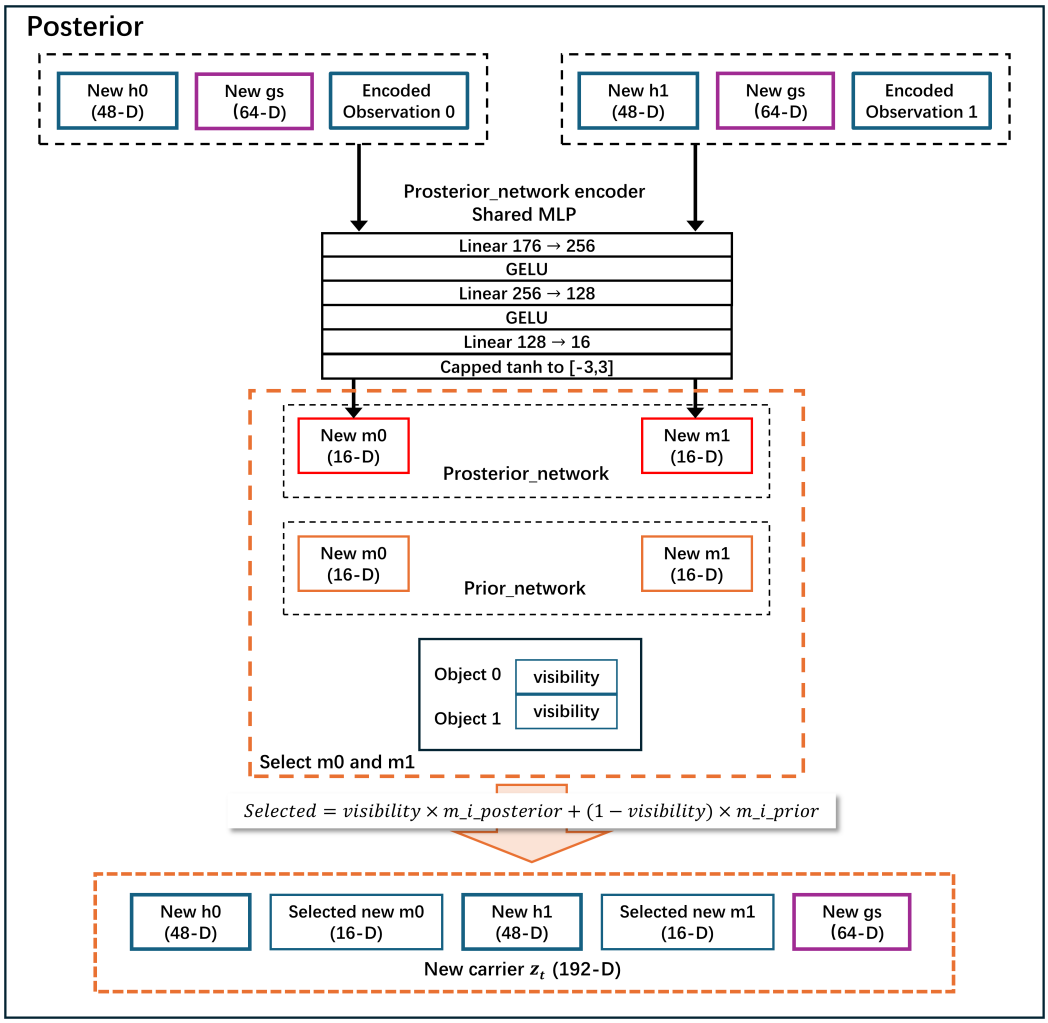


Figure B3: Posterior assimilation and carrier assembly. Visibility selects between prior and posterior middle components before the updated object states, selected middle components, and updated global state are packed in the fixed 48 + 16 + 48 + 16 + 64 layout.

The selected components are packed in fixed order as $\mathbf{z_t} = [\mathbf{h'_0}, \mathbf{m'_0}, \mathbf{h'_1}, \mathbf{m'_1}, \mathbf{gs'}] \in \mathbb{R}^{192}$. The update is deterministic; there is no sampling or reparameterization step.

In the selector equation, the binary visibility variable chooses the posterior middle component when the object is visible and the prior middle component otherwise. The assembled carrier is ordered as the updated object-0 recurrent state, selected object-0 middle component, updated object-1 recurrent state, selected object-1 middle component, and updated global recurrent state, with widths 48 + 16 + 48 + 16 + 64 = 192.

At the anchor, the stored 192-D carrier is post-recurrent-update and post-observation-assimilation. The one-shot intervention is applied to this complete carrier before autonomous rollout.

B.4 Training data and optimization

All data are generated with the deterministic simulator in Appendix B.1; no external physics dataset or pretrained trajectory corpus is used. The 5120 paired units share pre-anchor physical history and realized observation noise; the registered velocity edit is applied at the selected anchor and the simulator is rerun from the edited state. Dataset-generation and split seeds are 290413 and 290414.

Training combines rollout MSE with anchor-state, filtering-state, relative-state, and delta-velocity losses with weights 1.0/0.2/0.2/0.25/0.25; the recorded delta-velocity q90 scale is 1.2762992187. Contact/event labels, carrier/rank supervision, and intervention-method hyperparameter search are not used. Detailed training information can be found in the following table.

Table B4: Training data and optimization settings

Training item	**Value**
Dataset units	5120 total units; 10240 factual/counterfactual branches
Training split	4096 units
Development-validation split	512 units
Model-adequacy nonfinal split	512 units
Training anchors	t = 5, 6, 7, 8, 9
Batch construction	96 units / 192 branches; factual weight 0.5, counterfactual weight 0.5
Training duration	70 epoch-equivalents; 24 batches per epoch; 1,680 optimizer updates
Optimizer	AdamW; learning rate 0.0015; weight decay 0.00001; gradient clip norm 5.0
Scheduler	Cosine annealing over 70 epoch-equivalents; minimum learning rate 0.00015
Checkpoint selection	terminal model state after 1,680 updates; no best-validation selection loop

B.5 CHECKPOINT ROLES

All six checkpoints use the same architecture and training budget. Development checkpoints are used for method development and rank selection, whereas fresh checkpoints are reserved for frozen-method replication. Each checkpoint fits its own carrier and addressable map; numerical basis coordinates are not aligned across checkpoints. Detailed information on checkpoints can be found in the following talbe.

Table B5: Development and fresh-replication checkpoint roles

Public role	**Internal seed**	**Use in this paper**
Development model 1	291101	method development and rank selection
Development model 2	291102	method development and rank selection
Development model 3	291103	method development and rank selection
Fresh replication model 1	291301	frozen-method replication
Fresh replication model 2	291302	frozen-method replication
Fresh replication model 3	291303	frozen-method replication

B.6 MATCHED TRAINING REGIMES FOR THE COMPOSITIONAL EXTENSION

For the Joint and mechanistic extension, MF0–MF2 keep architecture, optimizer, scheduler, update count, and total branch presentations fixed; only intervention-support composition changes (as shown in Table B6). The resulting MF0/MF1/MF2 comparisons are interpreted as descriptive family-level associations rather than a universal causal law of training. Each family receives 322560 branch presentations. In MF2, a Joint-counterfactual branch changes both velocity components of one object while leaving all positions and the other object's primitive anchor state unchanged directly; Appendix C defines the exact Joint semantics. These extension families do not reopen the original Single rank-selection decision.

Table B6: Matched training regimes used in the compositional extension

Family	Per-update branch composition	Role in the paper	Evidence role
MF0: factual-only	192 factual branches	Contextual comparison for native Joint behavior and hidden-response organization.	Contextual model-family comparison
MF1: factual + Single-CF support	96 factual + 96 Single-counterfactual branches	Primary Single-support family; also serves as the baseline for the compositional extension.	Primary confirmatory family + extension baseline
MF2: factual + Single+Joint-CF support	96 factual + 48 Single-counterfactual + 48 Joint-counterfactual branches	Tests how explicit Joint support is associated with Joint rollout fidelity, specificity, and hidden-response organization.	Replicated Joint-support extension

B.7 Dataset and analysis split summary

Table B7 summarizes the data roles referenced throughout the paper; confirmatory and TEST roles are evaluation-only and do not enter carrier fitting or method selection.

Table B7: Dataset and analysis split

Purpose	Units	Where used
Model-training dataset	5120 total	B.4
Training split	4096	model fitting
Development-validation	512	model development
Model-adequacy nonfinal	512	adequacy check
Single carrier-fit	1024 / checkpoint	carrier + affine map fitting
Main Single confirmatory	256 / checkpoint / stratum	final Single assay
Temporal B1 carrier fit	128 / anchor / stratum	temporal fit
Temporal B1 method dev	128 / anchor / stratum	temporal development
Temporal B1 confirmatory	128 / anchor / stratum	temporal evaluation
Joint FIT	411 / checkpoint	whitening/operator fit
Joint TEST	101 / checkpoint	composition/addressability/specificity

C Counterfactual Pair Construction and Intervention Semantics

This appendix specifies the pairing, anchor-edit semantics, simulator references, evaluation strata, and matched Joint construction.

C.1 Same-history pairing

Each factual/counterfactual pair shares the same pre-anchor physical trajectory and realized observation-noise sequence:

$$\mathbf{s}^{CF}_{0:t-1} = \mathbf{s}^{F}_{0:t-1} \qquad \epsilon^{CF}_{0:t} = \epsilon^{F}_{0:t} \tag{C1}$$

Here, $\mathbf{s}^{F}_{0:t-1}$ and $\mathbf{s}^{CF}_{0:t-1}$ are the factual and counterfactual pre-anchor physical-state sequences, while $\epsilon^{F}_{0:t}$ and $\epsilon^{CF}_{0:t}$ are the realized observation-noise sequences shared by the pair.

The counterfactual branch first differs at the registered anchor edit; the same pairing rule applies to Joint units.

C.2 PRIMARY SINGLE VELOCITY-EDIT FAMILY

The velocity request is represented by a four-slot vector $\mathbf{e}$ ordered as $(\Delta v_{x,A}, \Delta v_{y,A}, \Delta v_{x,B}, \Delta v_{y,B})$. A Single request has exactly one nonzero slot:

$$\mathbf{e} = [\Delta v_{x,A} \quad \Delta v_{y,A} \quad \Delta v_{x,B} \quad \Delta v_{y,B}]^T \qquad \| \mathbf{e} \|_0 = 1 \tag{C2}$$

Here, $\mathbf{e} = (\Delta v_{x,A}, \Delta v_{y,A}, \Delta v_{x,B}, \Delta v_{y,B})^\mathsf{T} \in \mathbb{R}^4$ is the addressable request vector, and $\| \mathbf{e} \|_0$ counts its nonzero entries; a Single request has exactly one.

The registered Single family uses the following signed increments:

$$\delta v_{abs} \in \{0.066,\ 0.132,\ 0.264\} \qquad \delta v \in \{-\delta v_{abs}, +\delta v_{abs}\} \tag{C3}$$

Here, δv_{abs} is the registered absolute edit magnitude and δv is its signed realization.

At the anchor, all positions and the other three velocity coordinates remain unchanged. Later positions, contacts, and velocities arise only through simulator evolution; the edited-anchor speed remains within the registered support (maximum recorded bound approximately 0.6711).

Joint uses the same four-slot request but changes both velocity coordinates of one object:

$$\| \mathbf{e} \|_0 = 2 \qquad supp\,(\mathbf{e}) = \{(j, x), (j, y)\} \qquad j \in \{A, B\} \tag{C2a}$$

Here, $supp\,(\mathbf{e})$ is the set of nonzero request slots. The two slots (j, x) and (j, y) must belong to the same object $j \in \{A, B\}$.

Joint requests use the same four-slot representation with both velocity slots of one object nonzero; their Euclidean request magnitude and edited-anchor speed remain within the registered Single support.

C.3 SIMULATOR TRUTH AND MODEL-NATIVE COUNTERFACTUAL

After the anchor edit, the deterministic simulator is propagated from the edited state to define counterfactual truth. G1 uses the ordinary model update on the edited anchor observation; G2 reconstructs the same native counterfactual hidden state by full-hidden injection.

$$\mathbf{z}_t^{G2} = \mathbf{z}_t^F + \left(\mathbf{z}_t^{CF} - \mathbf{z}_t^F\right) = \mathbf{z}_t^{CF} \tag{C4}$$

Here, $\mathbf{z}_t^F$ is the factual anchor state, $\mathbf{z}_t^{CF}$ is the native counterfactual state produced by the ordinary edited-observation update, and $\mathbf{z}_t^{G2}$ is the state produced by full-hidden injection. Equation (C4) shows that G2 reconstructs $\mathbf{z}_t^{CF}$ exactly before rollout.

G1 is the model-adequacy reference and G2 an interface-equivalence control; neither establishes low-rank capacity or addressability. The same definitions apply to Joint units.

C.4 EVALUATION STRATA

S1 and S2 are scored separately; success in S1 cannot compensate for failure in S2.

Table C1: Evaluation strata and their operational roles

Stratum	Operational role
S1: NO_CONTACT_WITHIN_HORIZON	Neither the factual nor the counterfactual trajectory contains a detected future contact step over the 12-transition horizon; used to test local kinematics and autonomous propagation without later contact.

Stratum	Operational role
S2: EDIT_AFFECTS_FUTURE_CONTACT	The factual trajectory contains a detected future contact step and the edit produces a sufficiently large post-anchor factual–counterfactual trajectory difference; used to test downstream coupled consequences of the edit. Counterfactual contact, a contact/noncontact switch, or a changed contact time is not required.

C.5 Anchor and rollout support

The main Single assay uses anchor t = 7 and 12 autonomous transitions. Temporal tests use t = 5 - 9, including direct transport of the exact t = 7 interface; Joint is evaluated at t = 7 with the same horizon. Claims remain restricted to the registered edit and edited-anchor speed support.

C.6 Position-edit diagnostic boundary

The position stress test is development-only: at t = 7, one x or y coordinate is shifted by 0.05, 0.10, or 0.20, simulator truth is regenerated, and the intended patch is compared with no-patch, random equal-norm, and wrong-object alternatives. It does not enter Single rank selection or fresh-checkpoint confirmation.

C.7 Same-object two-component Joint velocity extension

Joint composition uses identity-matched factual/S_x/S_y/Joint routes sharing base unit, edited object, anchor, pre-anchor history, and realized noise; the two Single requests sum to the Joint request. The 512 units split into 411 FIT and 101 TEST (Appendix H.1).

For MF1/MF2, the rank-4 basis and addressable map are fitted only on Single examples; Joint hidden differences and Joint map examples do not enter fitting. MF1 tests unseen-Joint generalization; MF2 evaluates the same construction with Joint-supported training. Single rank selection is not reopened.

D Low-Rank Carrier Construction

This appendix defines the checkpoint-specific Single-derived carrier and the privileged oracle projections used to test subspace capacity. Joint hidden differences never enter the carrier-basis fit.

D.1 Hidden-difference matrix

For Single carrier-fit unit i at anchor t, define the native counterfactual-minus-factual carrier difference at the same post-observation/pre-transition interface:

$$\Delta \mathbf{z}_i = \mathbf{z}_{t,i}^{CF} - \mathbf{z}_{t,i}^{F} \in \mathbb{R}^{192} \tag{D1}$$

Here, i indexes a Single carrier-fit unit, t is the anchor, and $\Delta \mathbf{z}_i \in \mathbb{R}^{192}$ is its native counterfactual-minus-factual hidden difference at the common post-observation, pre-transition interface.

Each checkpoint contributes $N = 1024$ Single fit units, stacked row-wise as

$$\mathbf{D} = \begin{bmatrix} (\Delta \mathbf{z}_1)^\mathsf{T} \\ \vdots \\ (\Delta \mathbf{z}_N)^\mathsf{T} \end{bmatrix} \in \mathbb{R}^{N \times 192} \qquad N = 1024 \tag{D2}$$

Here, $N = 1024$ is the number of carrier-fit units per checkpoint and $\mathbf{D} \in \mathbb{R}^{N \times 192}$ stacks one transposed hidden difference per row.

The fit matrix contains only raw, uncentered Single carrier-fit differences; test units and later development or confirmatory outcomes do not enter basis fitting.

D.2 UNCENTERED SINGULAR-VALUE DECOMPOSITION

We factor the raw difference matrix without subtracting its mean:

$$\mathbf{D} = \mathbf{L}\,\mathbf{\Sigma}\,\mathbf{V}^{\mathsf{T}} \tag{D3}$$

Here, $\mathbf{L}$ and $\mathbf{V}$ contain the left and right singular vectors, and $\mathbf{\Sigma} = diag\,(\sigma_1, \ldots, \sigma_{192})$ contains the singular values. If $\mathbf{v}_k$ is the kth right singular vector, the candidate rank-r carrier basis is

$$\mathbf{U}_r = [\mathbf{v}_1, \ldots, \mathbf{v}_r] \in \mathbb{R}^{192\times r} \qquad \mathbf{U}_r^{\mathsf{T}}\mathbf{U}_r = \mathbf{I}_r \tag{D4}$$

Basis-vector signs are fixed only for deterministic artifact storage and have no scientific meaning.

D.3 PROJECTION AND ORACLE CAPACITY

The rank-r projector, privileged oracle coordinates, and projected reconstruction are

$$\begin{aligned} \mathbf{P}_r &= \mathbf{U}_r\mathbf{U}_r^{\mathsf{T}} \\ \mathbf{c}_i^{oracle} &= \mathbf{U}_r^{\mathsf{T}}\Delta\mathbf{z}_i \\ \widehat{\Delta\mathbf{z}}_i^{\,oracle} &= \mathbf{U}_r\mathbf{c}_i^{oracle} = \mathbf{P}_r\Delta\mathbf{z}_i \end{aligned} \tag{D5}$$

Here, $\mathbf{P}_r = \mathbf{U}_r\mathbf{U}_r^{\mathsf{T}}$ is the rank-r orthogonal projector, $\mathbf{c}_i^{oracle} = \mathbf{U}_r^{\mathsf{T}}\Delta\mathbf{z}_i$ is the privileged coefficient vector, and $\widehat{\Delta\mathbf{z}}_i^{\,oracle} = \mathbf{P}_r\Delta\mathbf{z}_i$ is the closest Euclidean reconstruction inside the carrier.

Because the oracle coordinates use the same unit's native counterfactual hidden state, the projected reconstruction tests carrier capacity rather than test-time addressability.

For Joint unit i, let $\Delta\mathbf{z}_i^J = \mathbf{z}_{t,i}^J - \mathbf{z}_{t,i}^F$. With the Single-derived $\mathbf{U}_4$ frozen,

$$\begin{aligned} \mathbf{c}_i^{J,U_4\text{-oracle}} &= \mathbf{U}_4^{\mathsf{T}}\Delta\mathbf{z}_i^J \\ \widehat{\Delta\mathbf{z}}_i^{\,J,U_4} &= \mathbf{P}_4\Delta\mathbf{z}_i^J \end{aligned} \tag{D5a}$$

Here, $\Delta\mathbf{z}_i^J = \mathbf{z}_{t,i}^J - \mathbf{z}_{t,i}^F$ is the native Joint hidden difference, while $\mathbf{U}_4$ and $\mathbf{P}_4$ remain the frozen Single-derived rank-4 basis and projector. These quantities are oracle references, not inputs to the addressable Joint route.

The Joint projection is likewise a privileged capacity reference and is not used to generate the addressable Joint patch.

D.4 DESCRIPTIVE SECOND-MOMENT CAPTURE

As a descriptive geometry summary, we report the cumulative squared-singular-value fraction

$$\rho_r = \frac{\sum_{k=1}^{r}\sigma_k^2}{\sum_{k=1}^{192}\sigma_k^2} \tag{D6}$$

Here, σ_k is the kth singular value and ρ_r is the cumulative fraction of the raw uncentered second moment captured by the first r right-singular directions.

Because the SVD is uncentered, this quantity describes raw second-moment concentration rather than mean-centered variance. It is descriptive only: it is neither a physical-energy measure nor the rank-selection criterion. As shown in Figure D1, the first few singular directions account for a large fraction of this raw second moment across all three development checkpoints: the cumulative ratio at rank 4 is 0.824, 0.802, and 0.866, rising to approximately 0.93 – 0.97 by rank 8. The vertical rank-4 marker indicates the rank selected by the registered rollout-and-control rule, rather than a spectral cutoff. The corresponding carrier spectrum is supplementary rank context.

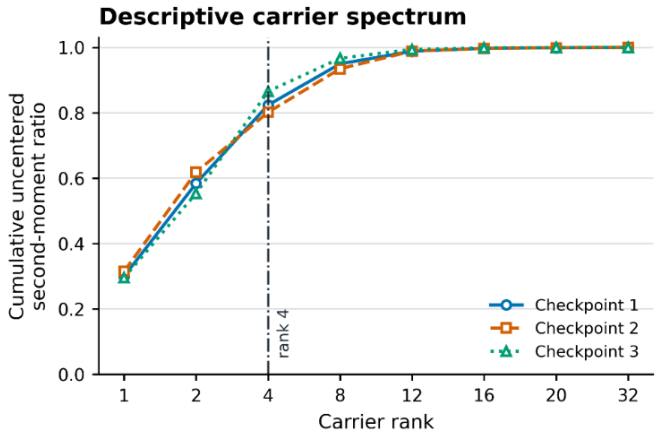


Figure D1: Descriptive carrier spectrum for the primary Single experiment. Curves show the cumulative uncentered second-moment ratio ρ_r across the three development checkpoints; the vertical line marks rank 4. Rank was selected by the registered rollout-and-control rule, not by a spectral threshold.

D.5 Registered rank sets

$$\mathcal{R} = \{1, 2, 4, 8, 12, 16, 20, 32\} \qquad \mathcal{R}_{diag} = \{64\} \qquad \mathcal{R}_{ref} = \{0, 192\} \tag{D7}$$

Here, $\mathcal{R}$ is the selectable rank grid, $\mathcal{R}_{diag}$ contains the diagnostic rank, and $\mathcal{R}_{ref}$ contains the no-patch and full-hidden reference ranks.

The selectable grid is fixed before development outcomes are read. Rank 64 is diagnostic; ranks 0 and 192 are no-patch and full-hidden references. The reported value is therefore the smallest passing tested rank on the registered grid, not a claim about untested ranks or nonlinear carriers.

D.6 Checkpoint and split boundaries

Each checkpoint fits its own basis from its 1024-unit carrier-fit split; no cross-checkpoint axis alignment is assumed. Development and confirmatory outcomes do not refit the basis, and fresh checkpoints use the already frozen construction.

The Joint extension reuses the rank selected by the primary Single development panel and does not reopen rank selection.

D.7 Single-derived rank-4 carrier in the Joint extension

For MF1 and MF2, the rank-4 carrier remains checkpoint-specific and Single-derived; Joint hidden differences are excluded from basis fitting.

The frozen rank-4 carrier has two distinct roles: Eq. (D5a) supplies a privileged Joint-capacity projection, while Appendix E uses the same Single-derived carrier with the Single-only addressable map to generate the actual Joint patch.

MF0 enters native Joint behavior and composition comparisons but not the Single-derived carrier/addressability analysis. The factual-whitened projectors in Appendix I are diagnostic coordinate transforms only and do not refit the carrier or alter the rollout patch.

E ADDRESSABLE COEFFICIENT MAP

This appendix defines the Single-only addressable map that converts a factual primitive state and requested velocity edit into coordinates in the checkpoint-specific carrier from Appendix D. The map is fitted only on Single carrier-fit examples and never reads the evaluation unit's native counterfactual hidden state.

E.1 INPUT ENCODING

$$\mathbf{x}_i = \begin{bmatrix} \mathbf{e}_i \\ \mathbf{s}^F_{t,i} \end{bmatrix} \in \mathbb{R}^{12} \qquad \mathbf{e}_i \in \mathbb{R}^4, \qquad \mathbf{s}^F_{t,i} \in \mathbb{R}^8 \tag{E1}$$

Here, $\mathbf{x}_i \in \mathbb{R}^{12}$ concatenates the four-slot edit request $\mathbf{e}_i \in \mathbb{R}^4$ and factual primitive state $\mathbf{s}^F_{t,i} \in \mathbb{R}^8$. Single fitting uses one nonzero edit slot; Joint evaluation uses the two velocity slots of the same object. No Joint-specific feature is added.

The four edit slots are the horizontal and vertical velocity changes for objects A and B. Single fitting uses one nonzero slot; Joint evaluation uses the two same-object velocity slots. The factual state contributes the eight primitive position/velocity coordinates, with no Joint-specific feature.

E.2 FIT-ONLY STANDARDIZATION

Each non-intercept feature is standardized with carrier-fit statistics only:

$$\widetilde{x}_{i,k} = \frac{x_{i,k} - \mu_k}{\max(\sigma_k, 10^{-6})} \qquad k = 1, \ldots, 12 \tag{E2}$$

Here, $x_{i,k}$ is feature k of unit i, μ_k and σ_k are its carrier-fit mean and standard deviation, and $\widetilde{x}_{i,k}$ is the standardized value. The floor 10^{-6} prevents division by a negligible scale.

Here μ_k and σ_k are the fit-split mean and standard deviation. The floor 10^{-6} avoids division by a negligible scale. These statistics are frozen for evaluation and are not recomputed in the temporal B2 transport test.

E.3 COEFFICIENT TARGETS

Single carrier-fit units provide privileged coefficient targets by projection onto the checkpoint-specific carrier:

$$\mathbf{c}_i^{oracle} = \mathbf{U}_r^{\mathsf{T}} \Delta \mathbf{z}_i \qquad \mathbf{C}_r = \mathbf{D}\mathbf{U}_r \in \mathbb{R}^{N \times r} \tag{E3}$$

Here, $\mathbf{c}_i^{oracle} \in \mathbb{R}^r$ is the privileged target for fit unit i, and $\mathbf{C}_r = \mathbf{D}\mathbf{U}_r \in \mathbb{R}^{N \times r}$ stacks these targets row-wise. Joint hidden differences are excluded from $\mathbf{C}_r$.

Each row is the privileged carrier-coordinate target for one Single fit unit, obtained by projecting its native hidden difference onto the checkpoint-specific carrier from Appendix D. Joint hidden differences are excluded from mapper targets.

E.4 RIDGE-AFFINE FIT

$$(\mathbf{W}_r, \mathbf{b}_r) = arg\,min_{\mathbf{W},\mathbf{b}} \sum_{i=1}^{N} \parallel \mathbf{c}_i^{oracle} - \left(\mathbf{W}^{\mathsf{T}} \widetilde{\mathbf{x}}_i + \mathbf{b}\right) \parallel_2^2 + \lambda \parallel \mathbf{W} \parallel_F^2 \qquad \lambda = 10^{-4} \tag{E4}$$

Here, $\mathbf{W}_r \in \mathbb{R}^{12\times r}$ is the ridge weight matrix, $\mathbf{b}_r \in \mathbb{R}^r$ is the affine intercept, and $\lambda = 10^{-4}$ is the fixed regularization coefficient. $\| \mathbf{W} \|_F$ is the Frobenius norm. The symbol $\mathbf{W}_r$ is unrelated to the whitening matrix $\mathbf{W}_{wh}$ in Appendix H.

Equation (E4) fits the weight matrix and intercept of the rank-specific addressable map defined in Eq. (E5). The ridge penalty applies to the weight matrix but not the intercept; the mapper family, regularization, and fit protocol are fixed from the Single procedure, with no Joint example or outcome-based alternative-map selection.

E.5 Test-time inference and patch

$$\widehat{\mathbf{c}} = f_{addr,r}\left(\mathbf{s}_t^F, \mathbf{e}\right) = \mathbf{W}_r^{\mathsf{T}}\widetilde{\mathbf{x}} + \mathbf{b}_r \tag{E5}$$

Here, $\widetilde{\mathbf{x}}$ is formed using the frozen carrier-fit statistics, and $\widehat{\mathbf{c}} \in \mathbb{R}^r$ is the test-time coefficient predicted by $f_{addr,r}$.

At test time the map receives only the factual primitive state, requested edit, and frozen fit statistics; it cannot access the native counterfactual hidden state, simulator future, future observations, or oracle coefficients.

$$\mathbf{z}_t^* = \mathbf{z}_t^F + \beta \mathbf{U}_r \widehat{\mathbf{c}} \qquad \beta = 1 \tag{E6}$$

Here, $\mathbf{z}_t^F$ is the factual anchor state, $\beta = 1$ is the frozen intervention strength, and $\mathbf{z}_t^\star$ is the once-patched hidden state released for autonomous rollout.

The predicted patch is applied once at the anchor; the addressable map is not consulted again during autonomous rollout. This separation is the operational meaning of addressability used in the paper.

E.6 Oracle capacity versus addressability

Table E1: Oracle-capacity and addressable coefficient routes

Route	Coefficient source	Allowed interpretation
Oracle capacity	Same-unit native hidden difference projected onto the rank-r carrier basis.	Tests whether the rank-r subspace can carry the required hidden difference; privileged upper bound only.
Addressable coefficient map	Predicted from the factual primitive state and the registered edit.	Primary intervention route; tests whether the carrier can be reached without a test-time native counterfactual hidden state.

E.7 Single-only affine composition for Joint requests

For Joint requests, the same four-slot representation and the same checkpoint-specific affine map fitted from Single examples are used unchanged. No Joint example enters standardization, coefficient fitting, regularization, or mapper selection.

$$\begin{gathered} \mathbf{e}_{xy} = \mathbf{e}_x + \mathbf{e}_y \\ \mathbf{c}_x = f_{addr,4}\left(\mathbf{s}_t^F, \mathbf{e}_x\right) \qquad \mathbf{c}_y = f_{addr,4}\left(\mathbf{s}_t^F, \mathbf{e}_y\right) \\ \mathbf{c}_0 = f_{addr,4}\left(\mathbf{s}_t^F, \mathbf{0}\right) \\ \mathbf{c}_{xy} = \mathbf{c}_x + \mathbf{c}_y - \mathbf{c}_0 \end{gathered} \tag{E7}$$

Here, $\mathbf{e}_x$ and $\mathbf{e}_y$ are the two matched Single requests, $\mathbf{e}_{xy} = \mathbf{e}_x + \mathbf{e}_y$ is the Joint request, $\mathbf{c}_x$ and $\mathbf{c}_y$ are their affine predictions, and $\mathbf{c}_0$ is the zero-edit baseline. The actual Joint rollout uses the direct full prediction $f_{addr,4}\left(\mathbf{s}_t^F, \mathbf{e}_{xy}\right)$, which equals the stated affine composition up to numerical precision.

Because the standardized map is affine and the factual state is fixed, the direct Joint coefficient and the component-wise affine composition agree up to numerical precision. This is an identity of the mapper only and does not imply additive native Joint hidden responses or dynamics-effective control.

The actual Joint rollout always uses the full affine prediction. The baseline and request-only components are introduced only for the patch-geometry diagnostic in Appendix I; rollout behavior and matched controls determine whether the generated Joint patch is dynamics-effective.

F ROLLOUT METRICS, ELIGIBILITY GATES, AND REGISTERED THRESHOLDS

This appendix defines the exact evaluation kernel used by the primary Single and temporal assays. The Joint extension reuses the same normalization, autonomous-rollout structure, physical-validity checks, and metric roles, with two-target scoring and matched specificity defined in F.10. M1–M5 are scientific metrics, M6E is an engineering eligibility gate, and M7 records rollout structure.

F.1 EVALUATION NOTATION AND FROZEN NORMALIZATION SCALES

For evaluation unit i, let t_i be the intervention anchor and $H = 12$ the number of autonomous future transitions. The denormalized decoded state at relative index j is $\widehat{\mathbf{X}}_{i,j} \in \mathbb{R}^{2\times 4}$, with object rows $o \in \{0, 1\}$ and columns (x, y, v_x, v_y). The anchor is $j = 0$ and the released future is $j = 1, \ldots, H$. The route-specific reference is $\mathbf{X}^{ref}_{i,j}$: counterfactual truth for patched, native-CF, wrong-object, and G2 routes, and factual truth for the factual/no-edit route.

The edited object is $o_i \in \{0, 1\}$ and the velocity-axis selector is $a_i \in \{0, 1\}$, so the edited primitive column is $c_i = 2 + a_i$. Primitive errors are normalized coordinatewise by the frozen scale matrix $\mathbf{s}^P$.

Table F1: Frozen primitive-coordinate normalization scales $\mathbf{s}^P$

	$\mathbf{x}$	$\mathbf{y}$	$\mathbf{v_x}$	$\mathbf{v_y}$
Object 0 (A)	0.64721	0.62436	0.40208	0.39198
Object 1 (B)	0.68680	0.55883	0.40071	0.35781

M2 uses the 13-dimensional derived representation $\phi(\mathbf{X})$ listed in Table F2 and its frozen scale vector $\mathbf{s}^D$.

Table F2: Frozen scales for the 13-dimensional derived representation

d	**Derived feature ϕ_d**	**Frozen scale $\mathbf{s}^D_d$**
1	relative_position_x	1.04125
2	relative_position_y	0.87639
3	relative_velocity_x	0.74275
4	relative_velocity_y	0.79797
5	center_position_x	0.32511
6	center_position_y	0.32665
7	center_velocity_x	0.06288
8	center_velocity_y	0.07177
9	summed_velocity_x	0.12576
10	summed_velocity_y	0.14354
11	0.5 x sum of squared speeds	0.06497
12	inter-object distance	0.35839
13	$\mathbf{1}\{distance \le 0.50000001\}$	1.00000

The scale registry was estimated once from 128 non-final threshold-calibration counterfactual simulator-truth units generated under the registered Single-edit protocol in Appendix C. These units obeyed the frozen support and feasibility rules and were not selected using model, low-rank, or final-test outcomes; the resulting scales were then frozen for all subsequent evaluation.

For continuous coordinates, the stored scale estimator is $\max(IQR/1.349,\ 1.4826\,MAD,\ 0.01)$; the contact-indicator scale is 1. These frozen scales are reused by all evaluators. M1–M4 add no clipping, epsilon, intervention-amplitude normalization, or metric-specific weighting.

F.2 ANCHOR FIDELITY AND FUTURE ROLLOUT METRICS (M1-M4)

M1 is the normalized absolute error of the directly edited velocity coordinate at the decoded anchor:

$$M_{1,i} = \frac{\left|\widehat{\mathbf{X}}_{i,0,o_i,c_i} - \mathbf{X}^{ref}_{i,0,o_i,c_i}\right|}{s^P_{o_i,c_i}} \qquad c_i = 2 + a_i \tag{F1}$$

Here, i indexes an evaluation unit, o_i is the edited object, $a_i \in \{0, 1\}$ selects horizontal or vertical velocity, $c_i = 2 + a_i$ is the corresponding primitive column, and $s^P_{o_i,c_i}$ is the frozen primitive-coordinate scale.

M2 is the unweighted RMS error of the 13 normalized derived quantities at the anchor:

$$M_{2,i} = \sqrt{\frac{1}{13}\sum_{d=1}^{13}\left(\frac{\phi_d\left(\widehat{\mathbf{X}}_{i,0}\right) - \phi_d\left(\mathbf{X}^{ref}_{i,0}\right)}{s^D_d}\right)^2} \tag{F2}$$

Here, $\phi_d(\cdot)$ is derived feature $d \in \{1, \ldots, 13\}$ from Table F2 and s^D_d is its frozen normalization scale.

M3 is the normalized RMSE over all 12 released future states, both objects, and all four primitive fields; the anchor is excluded:

$$M_{3,i} = \sqrt{\frac{1}{8H}\sum_{j=1}^{H}\sum_{o=0}^{1}\sum_{c=0}^{3}\left(\frac{\widehat{\mathbf{X}}_{i,j,o,c} - \mathbf{X}^{ref}_{i,j,o,c}}{s^P_{o,c}}\right)^2} \qquad H = 12 \tag{F3}$$

Here, $j = 1, \ldots, H$ indexes released future states, $o \in \{0, 1\}$ indexes objects, $c \in \{0, 1, 2, 3\}$ indexes primitive fields, and $H = 12$; the anchor $j = 0$ is excluded.

M4 is the maximum normalized anchor error over primitive coordinates declared unaffected by the edit:

$$M_{4,i} = \max_{(o,c):\, U_{i,o,c}=1} \frac{\left|\widehat{\mathbf{X}}_{i,0,o,c} - \mathbf{X}^{ref}_{i,0,o,c}\right|}{s^P_{o,c}} \tag{F4}$$

Here, $U_{i,o,c} \in \{0, 1\}$ is the unaffected-coordinate mask. The maximum is taken only over primitive slots declared unaffected by the requested edit.

The unaffected-coordinate mask contains seven primitive slots for Single edits; Joint scoring is defined separately in F.10.

F.3 POST-EDIT LAW AND GEOMETRY CONSISTENCY (M5)

M5 is the maximum of five checks: exact one-step simulator replay, circle overlap, arena-boundary violation, excessive per-step displacement, and consistency between the recorded anchor difference and the declared edit.

$$R^{dyn}_{i,j} = \sqrt{\frac{1}{8}\sum_{o=0}^{1}\sum_{c=0}^{3}\left(\frac{\widehat{\mathbf{X}}_{i,j+1,o,c} - \left[\mathcal{S}\left(\widehat{\mathbf{X}}_{i,j}, \mathbf{a}_{i,t_i+j}\right)\right]_{1,o,c}}{s^P_{o,c}}\right)^2} \qquad 0 \le j < H \tag{F5}$$

$$R^{overlap}_i = \frac{1}{0.01}\max_{0\le j\le H}\max\left(0.5 - \parallel \widehat{\mathbf{p}}_{i,j,1} - \widehat{\mathbf{p}}_{i,j,0} \parallel_2,\, 0\right) \tag{F6}$$

$$R_i^{arena} = \frac{1}{0.01} \max_{j,o,q \in \{x,y\}} \max\left(|\widehat{p}_{i,j,o,q}| - 4.75,\, 0\right) \tag{F7}$$

$$R_i^{step} = \frac{1}{0.01} \max_{\substack{0 \le j < H \\ o \in \{0,1\}}} \max\Big(\parallel \widehat{\mathbf{p}}_{i,j+1,o} - \widehat{\mathbf{p}}_{i,j,o} \parallel_2 - 0.25 \max\left(\parallel \widehat{\mathbf{v}}_{i,j,o} \parallel_2, \parallel \widehat{\mathbf{v}}_{i,j+1,o} \parallel_2\right) - 10^{-8},\, 0\Big) \tag{F8}$$

$$\mathbf{B}_i = \mathbf{X}_{i,0}^{CF} - \mathbf{X}_{i,0}^{F} \qquad R_i^{decl} = \frac{\parallel \mathbf{B}_i - \mathbf{D}_i \parallel_\infty}{10^{-8}} \tag{F9}$$

$$M_{5,i} = max\left\{ \max_{0 \le j < H} R_{i,j}^{dyn},\, R_i^{overlap},\, R_i^{arena},\, R_i^{step},\, R_i^{decl} \right\} \tag{F10}$$

In Eqs. (F5) – (F10), $\mathcal{S}$ is the exact one-step simulator, a_{i,t_i+j} is the recorded action, $\widehat{p}$ and $\widehat{v}$ are decoded positions and velocities, B_i is the recorded factual-to-counterfactual anchor difference, and D_i is the corresponding edit-declaration matrix. Specifically, D_i is zero at every unedited slot and equals the recorded anchor change in B_i at the declared edited slot. Thus, R_i^{decl} checks whether the recorded intervention matches its declaration; it does not depend on the model decode. The five terms entering M_5 measure simulator replay, overlap, arena-bound, step-size, and declaration consistency, respectively. M_5 is their maximum rather than a sum or weighted score. Collision and reflecting-wall consistency are captured through the exact replay check, and any nonfinite decoded output causes the law-residual check to fail.

The five M5 terms are not summed or weighted. Collision and reflecting-wall consistency enter through the exact replay, and nonfinite decoded outputs fail the law-residual path.

F.4 M6E NUMERICAL ELIGIBILITY

M6E is an engineering gate rather than a scientific metric. The main assay requires a finite final hidden state and finite decoded trajectory under loose explosion bounds; the temporal assay adds a per-decoded-state vector bound.

$$E_i^{main} \Leftrightarrow finite\left(\mathbf{z}_i^{final}\right) \wedge finite\left(\widehat{\mathbf{X}}_{i,0:H}\right) \wedge \parallel \mathbf{z}_i^{final} \parallel_2 \le 10^6 \wedge \parallel \widehat{\mathbf{X}}_{i,0:H} \parallel_\infty \le 10^6 \tag{F11}$$

The temporal assay additionally requires:

$$E_i^{temp} \Leftrightarrow E_i^{main} \wedge \max_{0 \le j \le H} \parallel vec\left(\widehat{\mathbf{X}}_{i,j}\right) \parallel_2 \le 10^6 \tag{F12}$$

In Eqs (F11) and (F12) $finite(\cdot)$ requires all entries to be finite, $\mathbf{z}_i^{final}$ is the final recurrent state, $vec(\cdot)$ flattens a decoded 2×4 state, and E_i^{main} and E_i^{temp} are the main and temporal numerical-eligibility predicates.

M6E does not separately test intermediate-hidden finiteness or generic execution/serialization failures. Failed or ineligible units remain in the denominator, and the same gate is applied to Joint reference, oracle, and addressable rollouts.

F.5 M7 AUTONOMOUS-ROLLOUT STRUCTURE

M7 records the required structure of every scored 12-transition autonomous rollout (Table F3).

Table F3: Required structural metadata for the autonomous rollout (M7)

M7 field	Required value
transition_count	12
decoder_count	13
observation_reads_after_anchor	0
teacher_forcing_steps	0
state_clamp_steps	0
future_oracle_reads	0
pass	True

The 13 decodes are the anchor plus 12 future states. Joint requests use the same rollout structure; only the one-shot anchor patch changes.

F.6 G2 FULL-HIDDEN INTERFACE-EQUIVALENCE CHECK

G2 is an interface-equivalence check: it adds the complete native counterfactual-minus-factual hidden difference to the factual anchor state, thereby reconstructing the native counterfactual hidden state used by G1, and then compares the decoded rollouts.

$$d_Y = \max_{i,j,o,c} \left| Y^{G2}_{i,j,o,c} - Y^{G1}_{i,j,o,c} \right| \qquad G2Pass \Leftrightarrow d_Y \leq 10^{-6} \tag{F13}$$

Here, Y^{G1} and Y^{G2} are the decoded native-counterfactual and full-hidden-injection rollouts, d_Y is their maximum absolute discrepancy over all indexed entries, and G2 passes when $d_Y \leq 10^{-6}$.

The hidden discrepancy is zero algebraically because the full-hidden injection reconstructs the same state released by G1. G2 therefore tests the injection interface, not low-rank capacity or addressability.

F.7 PROGRAMMATIC S1/S2 STRATA

The following detector is applied to released future states in the anchor-plus-future sequence; the anchor itself is excluded.

$$\Gamma(\mathbf{X}) = \left\{ j \in \{1, \ldots, H\} : \| \mathbf{p}_{j,1} - \mathbf{p}_{j,0} \|_2 \leq 0.500001 \vee \| \mathbf{V}_j - \mathbf{V}_{j-1} \|_F > 10^{-5} \right\} \tag{F14}$$

The detector applies only to released future states.

$$S1 \Leftrightarrow \Gamma\left(\mathbf{X}^F_{0:H}\right) = \varnothing \wedge \Gamma\left(\mathbf{X}^{CF}_{0:H}\right) = \varnothing \tag{F15}$$

$$S2 \Leftrightarrow \Gamma\left(\mathbf{X}^F_{0:H}\right) \neq \varnothing \wedge \max_{k > t_i, o, c} \left| X^{CF}_{k,o,c} - X^F_{k,o,c} \right| \geq 10^{-4} \tag{F16}$$

In Eqs. (F14) – (F16), $\Gamma(\mathbf{X})$ is the set of detected future contact-related steps, $\mathbf{p}_{j,o}$ is object o's position, and $\mathbf{V}_j$ is the two-object velocity matrix.

S1 isolates trajectories for which neither branch contains a detected contact step over the 12-transition evaluation window. The detector is applied only to future indices $j = 1, \ldots, H$, so the anchor itself is not classified as a contact step. A future step is flagged when either the two object centers come within the registered contact-distance tolerance or the joint velocity array exhibits a change larger than 10^{-5}. Thus, S1 requires both the factual and counterfactual contact-step lists to remain empty throughout the evaluated future.

S2 instead selects interaction-sensitive cases in which the factual trajectory contains at least one detected contact step and the requested velocity edit produces a nontrivial subsequent trajectory change. The latter condition requires the maximum absolute factual–counterfactual difference over the stored post-anchor primitive states to be at least 10^{-4}. S2 does not require the counterfactual

branch itself to make contact, nor does it require the edit to change contact/non-contact status or contact timing.

Accordingly, S1 and S2 distinguish a contact-free regime from a factual-contact regime in which the edit has a measurable downstream effect; they are evaluated separately rather than treated as complementary exhaustive partitions.

F.8 Registered thresholds and unit/cell decision rule

The main low-rank and temporal assays use the same frozen threshold registry (Table F4). Comparisons are inclusive and nonfinite unit metrics fail.

Table F4: Registered M1-M5 thresholds for S1 and S2

Metric	S1 threshold	S2 threshold
M1	0.38876	0.46698
M2	0.65309	0.65681
M3	0.28745	0.55140
M4	0.33938	0.25616
M5	20.26019	10.47116

For each unit and stratum, the scientific condition is the joint M1–M5 threshold test, the eligibility condition combines the numerical/structural gates, and the final unit pass requires both:

$$Q_i \Leftrightarrow \bigwedge_{k=1}^{5} \left[finite\left(M_{i,k}\right) \wedge M_{i,k} \leq \theta_{s,k} \right] \qquad E_i \Leftrightarrow M6E_i \wedge M7_i \qquad J_i \Leftrightarrow Q_i \wedge E_i \tag{F17}$$

$$Coverage = \frac{N_{joint\text{-}pass}}{N_{expected}} \tag{F18}$$

The cell size is 256 units per checkpoint $\times$ stratum in the main confirmatory assay and 128 per checkpoint $\times$ anchor $\times$ stratum in the temporal assay. Failed or ineligible units remain in the denominator.

$$CellPass \Leftrightarrow CountOK \wedge EligOK \wedge CoverageOK \wedge MedianOK \wedge ControlsPass \tag{F19}$$

In Eqs. (F17) – (F19), Q_i is the joint M1–M5 scientific predicate, E_i is the numerical/structural eligibility predicate, J_i is the unit-pass predicate, $\theta_{s,k}$ is the frozen threshold for metric k in stratum s, and Coverage uses all expected units, including failures and ineligible units, in the denominator.

A cell passes only when all required eligibility/structure checks hold, unit-pass coverage is at least 0.8, every M1–M5 cell median is finite and within its stratum threshold, and the registered control condition in Appendix G passes. S1 and S2 are never pooled before this decision.

F.9 Checkpoint and panel aggregation

A checkpoint passes only if both strata pass:

$$SeedPass_q = CellPass_{q,S1} \wedge CellPass_{q,S2} \tag{F20}$$

The registered three-checkpoint panel requires at least two passing checkpoints:

$$PanelPass \Leftrightarrow \left| \{ q \in \{1,2,3\} : SeedPass_q \} \right| \geq 2 \tag{F21}$$

The temporal tests apply the same rule independently at each anchor or transport target:

$$\begin{aligned} ExistenceOverall &\Leftrightarrow \bigwedge_{t\in\{5,6,7,8,9\}} PanelPass_t \\ TransportOverall &\Leftrightarrow \bigwedge_{t\in\{5,6,8,9\}} PanelPass_t \end{aligned} \tag{F22}$$

Here, $q \in \{1, 2, 3\}$ indexes independently trained checkpoints, $SeedPass_q$ requires both S1 and S2 cells to pass, and $PanelPass$ requires at least two passing checkpoints. The two conjunctions require this panel rule at every registered B1 anchor or B2 transport target.

Trajectory units are not treated as independent model-level replications. Appendix G supplies the negative-control requirements, and the Joint extension does not alter the preregistered Single panel rule.

F.10 Same-object Joint scoring and specificity reporting

For Joint edits, both directly edited velocity components use the M1 primitive-coordinate normalization; their maximum is the Joint direct-target summary. M2 and M3 retain their original roles, the unaffected mask has six entries, and the M5 declaration check permits exactly the two declared velocity changes.

For $a \in \{0, 1\}$ indexing horizontal and vertical velocity axes,

$$M_{1,i}^{Joint} = \max_{a\in\{0,1\}} \frac{\left|\widehat{\mathbf{X}}_{i,0,o_i,2+a} - \mathbf{X}^{ref}_{i,0,o_i,2+a}\right|}{s^{P}_{o_i,2+a}} \tag{F23}$$

Here, $a = 0$ and $a = 1$ index the two velocity axes of edited object o_i. The Joint M1 score is the larger of the two normalized direct-target errors.

Native Joint, the privileged oracle route, and Addressable Joint are evaluated on the same 101 TEST units for each checkpoint. Each intended Joint request is matched to no-patch/sham, random equal-norm, wrong-object, and wrong-vector controls; Appendix H gives the aggregation procedure. These Joint analyses are separate from the preregistered Single evaluation and do not change its thresholds or fresh-checkpoint replication results.

G Control Routes and Temporal Generalization Protocols

This appendix defines the reference and control routes used to establish target specificity, the B1/B2 temporal-generalization protocols, the matched Joint controls, and the position-edit negative contrast. Exact metric and panel rules are in Appendix F.

G.1 Positive and structural reference routes

All reported actions are zero, so action shuffling is not a scientific contrast. Table G1 supplies the common route vocabulary: G1 is the native-counterfactual adequacy reference, G2 checks hidden-injection equivalence, and the remaining controls test no-op behavior, perturbation specificity, timing, leakage, and identity/order. G1/G2 are prerequisites rather than substitutes for negative controls.

Table G1: Reference and control routes used in the intervention assay

Route / control	Purpose
G0_FACTUAL_NO_EDIT	Checks ordinary factual rollout from the same model and interface.
SIMULATOR_COUNTERFACTUAL_TRUTH	External physical reference generated from the edited simulator state.
G1_NATIVE_COUNTERFACTUAL_FULL_HIDDEN	Checks whether the model can natively represent and propagate the complete edited current state.

Route / control	Purpose
G2_FULL_HIDDEN_EQUIVALENCE	Checks that hidden injection reproduces the native counterfactual route to the registered numerical tolerance.
SHAM_EDIT	Exercises the edit/map path with zero physical change; should behave as a no-op.
RANK_ZERO	Explicit no-patch reference for the low-rank intervention family.
RANDOM_ORTHOGONAL_EQUAL_NORM	Tests whether an equal-size but unrelated hidden perturbation can mimic the targeted result.
WRONG_OBJECT_EDIT	Tests object specificity by routing the edit to the other object while keeping the edit family comparable.
WRONG_TIME	Tests timing specificity by applying the registered mechanism at an adjacent or otherwise registered incorrect anchor.
LEAKAGE_NEGATIVE_TEST	Verifies that building the intervention does not require counterfactual future payload.
IDENTITY_ORDER_CONTROL	Verifies object-slot identity and decoder/output ordering.

G.2 B1: B1: PER-ANCHOR RECOVERABILITY

B1 tests recoverability by refitting the already frozen Single rank-4 procedure independently at $t = 5, 6, 7, 8, 9$. For each checkpoint and anchor, the carrier basis, feature standardization, and ridge-affine map use only that anchor's Single carrier-fit split; evaluation uses the corresponding sealed cells. B1 therefore tests whether the interface can be recovered at each anchor, not whether carrier coordinates are identical across time. Correlated data allocation can be found in Table G2.

Table G2: Temporal data allocation for the B1 per-anchor analysis

Temporal data role	Per anchor $\times$ stratum	Total across five anchors and two strata
Carrier fit	128 units	1280 units
Method development	128 units	1280 units
Confirmatory nonfinal	128 units	1280 units

G.3 B2: DIRECT REUSE OF THE FROZEN T = 7 INTERFACE

B2 tests direct reuse by transporting each checkpoint's exact $t = 7$ carrier basis, ridge weights/intercept, feature means/scales, rank 4, and intervention strength to $t = 5, 6, 8, 9$. No target-anchor refit, restandardization, rank change, or intervention-strength adjustment is allowed; B1 at the same target is the refitted reference.

Both B1 and B2 use the frozen M1-M5, M6E, M7, control, coverage, and 2-of-3 checkpoint rules in Appendix F; the temporal analysis does not change the registered cell or panel decision rule.

G.4 TEMPORAL RESULTS AND INTERPRETATION BOUNDARY

At every tested anchor $t = 5$ - 9, B1 satisfied the registered 2-of-3 checkpoint rule. B2 also satisfied the 2-of-3 rule at every transport target $t = 5, 6, 8, 9$ using the unchanged $t = 7$ interface. Figure G1 summarizes these checkpoint-level temporal results, showing successful B1 recovery across all tested anchors and successful B2 reuse of the fixed $t = 7$ interface at all four neighboring transport targets. These results establish local recoverability and direct reuse only over the tested nearby window; they do not imply globally time-invariant latent coordinates or temporal transport of the later Joint extension.

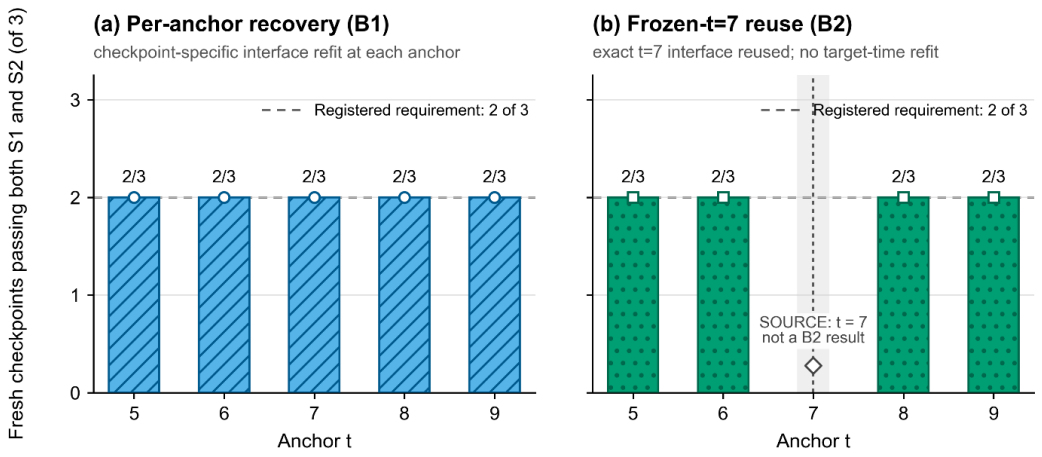


Figure G1: Temporal recoverability and direct reuse of the primary Single rank-4 interface. (a) B1 refits a checkpoint-specific interface at each anchor $t \in \{5, 6, 7, 8, 9\}$. (b) B2 transports the exact $t = 7$ interface to $t \in \{5, 6, 8, 9\}$ without refitting or restandardization; the $t = 7$ marker denotes the source interface.

G.5 Matched Joint controls for compositional addressability

Joint addressability is compared with four matched scientific controls: SHAM/NO_PATCH; RANDOM_ORTHOGONAL_EQUAL_NORM; WRONG_OBJECT_JOINT; and MATCHED_WRONG_VECTOR. The last keeps the edited object and request norm fixed while changing vector direction, testing specificity beyond perturbation magnitude.

Intended and control routes are identity-matched by base unit, anchor, edited object, and declared request and are scored on the same TEST units. Appendix H reports direct-target/future-error separation and the matched fraction for which each control is worse than the intended route. Leakage and identity/order remain integrity checks; these Joint controls do not affect Single rank selection or fresh-checkpoint replication.

G.6 Position-edit negative contrast

The development-only position diagnostic shifts one x or y coordinate of one object by 0.05, 0.10, or 0.20 at the anchor while holding all velocities and the remaining primitive anchor coordinates fixed; simulator counterfactual truth is regenerated from the edited state (as introduced in Appendix C.6). The intended low-rank patch is compared with rank-zero/no-patch, random equal-norm, and wrong-object alternatives.

Across all three amplitudes, the intended position patch, the rank-zero/no-patch route, and the random equal-norm control all passed the raw rollout gate in 3/3 development checkpoints. In contrast, wrong-object specificity and the full specificity criterion passed in 0/3. Figure G2a shows that raw rollout success is not specific to the intended patch, and Figure G2b shows that the specificity tests fail at every amplitude. The position edit therefore serves as a negative contrast: raw rollout improvement alone is not enough to identify a valid intervention. This does not mean that position information is absent or that no position intervention can work.

H Joint Composition and Addressability Protocols

This appendix defines the matched populations and metrics supporting the Joint addressability, composition, and specificity analyses in Sections 3.4 and 4.2–4.4. Single edits one velocity component of one object; Joint edits both velocity components of the same object at the same anchor.

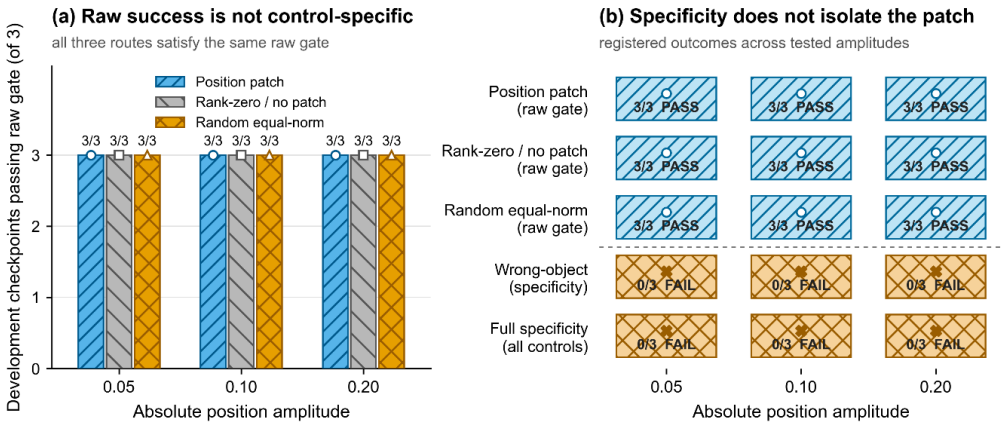


Figure G2: Position-edit negative contrast. (a) The intended position patch, rank-zero/no-patch route, and random equal-norm route all satisfy the raw rollout gate in 3/3 development checkpoints at every tested amplitude. (b) Wrong-object specificity and the full specificity criterion pass in 0/3 checkpoints at every amplitude, so the raw gate does not isolate the intended intervention.

H.1 Exact identity-matched quadruplets and split

For each base unit, the identity-matched quadruplet contains the factual route, horizontal Single route, vertical Single route, and Joint route. All four share unit identity, edited object, anchor, pre-anchor history, and realized observation noise. The two Single requests sum exactly to the Joint request.

$$\mathbf{e}_{xy,i} = \mathbf{e}_{x,i} + \mathbf{e}_{y,i}. \tag{H1}$$

Here, $\mathbf{e}_{x,i}$ and $\mathbf{e}_{y,i}$ are the matched horizontal and vertical Single requests for unit i, and $\mathbf{e}_{xy,i}$ is their same-object Joint request.

The edited object and Joint request vector are verified directly from the factual-to-Joint anchor-state difference. The 512 unique base units are split before analysis into 411 FIT and 101 TEST units; row order is never used as an identity key.

The 411 FIT units are used only for factual-reference whitening and local route-operator fitting; the 101 TEST units are reserved for composition, route comparison, patch geometry, and matched specificity.

H.2 FIT-only factual-reference whitening

Whitening is fitted checkpoint-wise from factual hidden states in the 411 FIT units only. The factual FIT mean, covariance, retained eigensystem, and frozen whitening matrix are defined in Eq. (H2).

$$\begin{aligned} \mu_F &= mean\left(\mathbf{Z}_F\right) \\ \mathbf{C}_F &= cov\left(\mathbf{Z}_F - \mu_F\right) = \mathbf{U}_{wh}\mathbf{\Lambda}_{wh}{\mathbf{U}_{wh}}^{\top} \\ \mathbf{W}_{wh} &= \mathbf{U}_{wh,keep}\mathbf{\Lambda}_{wh,keep}^{-1/2}{\mathbf{U}_{wh,keep}}^{\top} \end{aligned} \tag{H2}$$

Here, $\mathbf{Z}_F$ contains factual FIT hidden states, μ_F and $\mathbf{C}_F$ are their mean and covariance, $\mathbf{U}_{wh}$ contains covariance eigenvectors, $\mathbf{\Lambda}_{wh} = diag\left(\lambda_{wh,1}, \lambda_{wh,2}, \ldots\right)$ contains eigenvalues, and $\mathbf{W}_{wh}$ is the retained-eigensystem whitening matrix. The subscript 'keep' denotes directions satisfying the registered eigenvalue threshold.

An eigen-direction is retained when $\lambda_{wh,j}/\lambda_{wh,max} > 10^{-10}$. The subscript "keep" denotes the retained covariance eigenvectors and eigenvalues; hidden states are treated as column vectors.

$$wh\left(\mathbf{z}\right) = \mathbf{W}_{wh}\left(\mathbf{z} - \mu_F\right) \tag{H3}$$

Here, $wh\left(\mathbf{z}\right) = \mathbf{W}_{wh}\left(\mathbf{z} - \mu_F\right)$ maps a raw hidden state into the checkpoint-specific factual-whitened coordinates.

The TEST-time map in Eq. (H3) is a checkpoint-specific factual reference gauge only; it neither uses TEST units nor refits the Single-derived carrier or addressable intervention.

H.3 Static Joint composition

For each TEST unit, Eq. (H4) defines the factual-referenced whitened hidden responses for the matched horizontal Single, vertical Single, and Joint routes.

$$\begin{aligned}
\Delta\mathbf{z}_{x,i}^{wh} &= wh\left(\mathbf{z}_{S_x,i}\right) - wh\left(\mathbf{z}_{F,i}\right) \\
\Delta\mathbf{z}_{y,i}^{wh} &= wh\left(\mathbf{z}_{S_y,i}\right) - wh\left(\mathbf{z}_{F,i}\right) \\
\Delta\mathbf{z}_{xy,i}^{wh} &= wh\left(\mathbf{z}_{J_{xy},i}\right) - wh\left(\mathbf{z}_{F,i}\right) \\
\Delta\mathbf{z}_{add,i}^{wh} &= \Delta\mathbf{z}_{x,i}^{wh} + \Delta\mathbf{z}_{y,i}^{wh}
\end{aligned} \tag{H4}$$

Here, $\Delta\mathbf{z}_{x,i}^{wh}$ and $\Delta\mathbf{z}_{y,i}^{wh}$ are factual-referenced matched Single responses, $\Delta\mathbf{z}_{xy,i}^{wh}$ is the native Joint response, and $\Delta\mathbf{z}_{add,i}^{wh}$ is the sum of the two Single responses.

The x and y subscripts denote horizontal and vertical velocity edits. The primary static composition error in Eq. (H5) compares the native Joint response with the sum of the matched Single responses.

$$E_{add,i} = \frac{2 \parallel \Delta\mathbf{z}_{xy,i}^{wh} - \Delta\mathbf{z}_{add,i}^{wh} \parallel_2}{\parallel \Delta\mathbf{z}_{xy,i}^{wh} \parallel_2 + \parallel \Delta\mathbf{z}_{add,i}^{wh} \parallel_2 + \varepsilon}, \qquad \varepsilon = 10^{-12} \tag{H5}$$

Here, $E_{add,i}$ is the symmetric normalized distance between the native Joint response and the additive Single prediction, and $\varepsilon = 10^{-12}$ prevents division by zero.

Lower values of the error in Eq. (H5) indicate a more additive native Joint response; the response-norm and cosine quantities in Eq. (H6) are supporting diagnostics only.

$$\begin{aligned}
R_{norm,i} &= \frac{\parallel \Delta\mathbf{z}_{xy,i}^{wh} \parallel_2}{\parallel \Delta\mathbf{z}_{add,i}^{wh} \parallel_2 + \varepsilon} \\
r_{norm,i} &= \left|\log\left(\max\left(R_{norm,i}, \varepsilon\right)\right)\right| \\
\cos_{add,i} &= \frac{\left(\Delta\mathbf{z}_{xy,i}^{wh}\right)^\top \Delta\mathbf{z}_{add,i}^{wh}}{\parallel \Delta\mathbf{z}_{xy,i}^{wh} \parallel_2 \parallel \Delta\mathbf{z}_{add,i}^{wh} \parallel_2 + \varepsilon} \\
\cos_{xy,i} &= \frac{\left(\Delta\mathbf{z}_{x,i}^{wh}\right)^\top \Delta\mathbf{z}_{y,i}^{wh}}{\parallel \Delta\mathbf{z}_{x,i}^{wh} \parallel_2 \parallel \Delta\mathbf{z}_{y,i}^{wh} \parallel_2 + \varepsilon}
\end{aligned} \tag{H6}$$

Here, $R_{norm,i}$ compares response magnitudes, $r_{norm,i}$ is its absolute log deviation from one, $\cos_{add,i}$ compares Joint and additive directions, and $\cos_{xy,i}$ compares the two Single-response directions.

The static error in Eq. (H5) is the primary composition quantity used in the main analysis.

H.4 Dynamic composition under the actual recurrence

For each TEST unit, Eq. (H7) compares a synthetic additive initialization $\mathbf{z}_{add,i}^{(0)}$ with the native Joint initialization $\mathbf{z}_{joint,i}^{(0)}$.

$$\begin{aligned} \mathbf{z}_{add,i}^{(0)} &= \mathbf{z}_i^{S_x} + \mathbf{z}_i^{S_y} - \mathbf{z}_i^{F} \\ \mathbf{z}_{joint,i}^{(0)} &= \mathbf{z}_i^{Jxy} \end{aligned} \tag{H7}$$

Here, $\mathbf{z}_i^F$, $\mathbf{z}_i^{S_x}$, $\mathbf{z}_i^{S_y}$, and $\mathbf{z}_i^{Jxy}$, are the native 192-D post-anchor hidden states for the factual, matched horizontal Single, matched vertical Single, and same-object Joint routes, respectively.

Equation (H8) applies the frozen checkpoint's exact autonomous recurrent transition under the recorded zero action, with no future observation assimilation, state clamp, or hidden intervention, and evaluates the discrepancy at the registered horizons.

$$\begin{aligned} \mathbf{z}_{add,i}^{(k)} &= T^k\left(\mathbf{z}_{add,i}^{(0)}\right), \qquad \mathbf{z}_{joint,i}^{(k)} = T^k\left(\mathbf{z}_{joint,i}^{(0)}\right) \\ \mathbf{z}_{add,i}^{wh,(k)} &= wh\left(\mathbf{z}_{add,i}^{(k)}\right), \qquad \mathbf{z}_{joint,i}^{wh,(k)} = wh\left(\mathbf{z}_{joint,i}^{(k)}\right) \\ E_{dyn,i}(k) &= \frac{2 \parallel \mathbf{z}_{add,i}^{wh,(k)} - \mathbf{z}_{joint,i}^{wh,(k)} \parallel_2}{\parallel \mathbf{z}_{add,i}^{wh,(k)} \parallel_2 + \parallel \mathbf{z}_{joint,i}^{wh,(k)} \parallel_2 + \varepsilon} \end{aligned} \tag{H8}$$

Here, T^k denotes k applications of the checkpoint's exact recurrent transition, $\mathbf{z}_{add,i}^{wh,(k)}$ and $\mathbf{z}_{joint,i}^{wh,(k)}$ are their whitened states, and $E_{dyn,i}(k)$ is their symmetric normalized discrepancy after k transitions.

The dynamic discrepancy $E_{dyn,i}(k)$ defined in Eq. (H8) is smaller when the additive trajectory remains closer to the native Joint trajectory after k recurrent transitions. This diagnostic does not use the rank-4 carrier $\mathbf{U}_4$ or the fitted local operators in Appendix I.

H.5 Same-101 Native Joint, rank-4 oracle, and Addressable Joint comparison

For each MF1 and MF2 checkpoint, Native Joint, the privileged rank-4 oracle, and Addressable Joint are evaluated on the same 101 TEST units. Route-level coverage and rollout error are defined in Eq. (H9).

$$\begin{aligned} I_{r,i}^{pass} &= \mathbb{1}\left[JointPass(r,i)\right], \\ \pi_r &= \frac{1}{101}\sum_{i=1}^{101} I_{r,i}^{pass}, \\ e_{r,i,t} &= RMSE\left(\widehat{\mathbf{y}}_{r,i,t}, \mathbf{y}_{i,t}\right), \\ R_{r,t} &= median_i\, e_{r,i,t}. \end{aligned} \tag{H9}$$

Here, $r \in \{Native, Oracle, Addressable\}$ is a route label, not carrier rank; $I_{r,i}^{pass}$ is the unit-pass indicator, scalar π_r is route coverage, $e_{r,i,t}$ is primitive-state RMSE, and $R_{r,t}$ is its median over the 101 TEST units.

Checkpoint summaries are computed first and family summaries are then taken across the three checkpoints. The common MF1/MF2 TEST comparison contains 606 unique checkpoint-unit pairs. The pass-indicator notation is deliberately distinct from the carrier-projector notation in Appendix I.

H.6 Matched Joint specificity

For each metric (Joint M1, M3, or terminal RMSE), the intended addressable route is compared with each matched control on the same TEST units.

$$\begin{aligned} \Delta_m &= median_i m_{c,i} - median_i m_{a,i}, \\ w_m &= \frac{1}{101}\sum_{i=1}^{101} \mathbb{1}\left[m_{c,i} > m_{a,i}\right]. \end{aligned} \tag{H10}$$

Here, $m_{a,i}$ and $m_{c,i}$ are the intended-addressable and matched-control errors for metric m, Δ_m is the difference of their medians, and w_m is the fraction of matched units for which the control error is larger.

The separation is control error minus intended-route error; positive median separation means that the control is worse. The paired-worse fraction reports how often this ordering holds within matched units. Controls are no-patch/sham, random orthogonal equal-norm, wrong-object Joint, and matched wrong-vector; checkpoints remain the replication level.

H.7 Interpretation boundary

Static composition is the primary composition evidence and dynamic composition is supporting. The MF0/MF1/MF2 matched budgets are defined in Appendix B.6; their differences are interpreted as descriptive family-level associations, not universal causal effects of training.

I Carrier-Relative and Recurrent-Dynamics Diagnostics

This appendix gives the post-hoc diagnostics supporting the Section 4.4 interpretation of the rank-4 carrier as a compact intervention-entry interface rather than a closed four-dimensional dynamical state. All analyses are checkpoint-specific, use the FIT-only factual-reference whitening from Appendix H.2, and do not alter rank selection or confirmatory decisions.

I.1 Whitened $\mathbf{U_4}$ projector and complement

Let the checkpoint-specific Single-derived rank-4 carrier be expressed in raw hidden coordinates and use the FIT-only whitening matrix from Appendix H.2. We orthonormalize the whitened carrier and then define its projector and orthogonal complement:

$$\begin{aligned} \mathbf{Q}_U &= orth\left(\mathbf{W}_{wh}\mathbf{U}_4\right) \\ \mathbf{P} &= \mathbf{Q}_U\mathbf{Q}_U^\top \\ \mathbf{Q} &= \mathbf{I}_{192} - \mathbf{P} \end{aligned} \tag{I1}$$

Here, $orth(\cdot)$ returns an orthonormal basis for the columns of $\mathbf{W}_{wh}\mathbf{U}_4$, $\mathbf{Q}_U \in \mathbb{R}^{192\times 4}$ is that basis, $\mathbf{P} = \mathbf{Q}_U\mathbf{Q}_U^\top$ is the carrier projector, and $\mathbf{Q} = \mathbf{I}_{192} - \mathbf{P}$ is the complementary projector.

The projector in Eq. (I1) maps onto the four-dimensional carrier in factual-whitened coordinates, while the complementary projector maps onto the remaining 188 dimensions. These projectors are diagnostic only and do not refit the carrier.

I.2 Composition inside $\mathbf{U_4}$ and its complement

For each TEST unit, let the native Joint response and the additive combination of the matched Single responses be the whitened quantities defined in Appendix H.3; their difference is the composition defect.

$$\mathbf{d}_i = \Delta\mathbf{z}^{wh}_{xy,i} - \Delta\mathbf{z}^{wh}_{add,i}$$

Equation (I2) decomposes the native Joint response, additive response, and composition defect into carrier and orthogonal-complement components.

$$\begin{aligned} \mathbf{a}_{U,i} &= \mathbf{P}\Delta\mathbf{z}^{wh}_{xy,i} \qquad \mathbf{b}_{U,i} = \mathbf{P}\Delta\mathbf{z}^{wh}_{add,i} \qquad \mathbf{d}_{U,i} = \mathbf{P}\mathbf{d}_i \\ \mathbf{a}_{Q,i} &= \mathbf{Q}\Delta\mathbf{z}^{wh}_{xy,i} \qquad \mathbf{b}_{Q,i} = \mathbf{Q}\Delta\mathbf{z}^{wh}_{add,i} \qquad \mathbf{d}_{Q,i} = \mathbf{Q}\mathbf{d}_i \\ E_{add,U,i} &= \frac{2 \parallel \mathbf{d}_{U,i} \parallel_2}{\parallel \mathbf{a}_{U,i} \parallel_2 + \parallel \mathbf{b}_{U,i} \parallel_2 + \varepsilon} \\ E_{add,Q,i} &= \frac{2 \parallel \mathbf{d}_{Q,i} \parallel_2}{\parallel \mathbf{a}_{Q,i} \parallel_2 + \parallel \mathbf{b}_{Q,i} \parallel_2 + \varepsilon} \end{aligned} \tag{I2}$$

Here, $\mathbf{d}_i$ is the composition defect; $\mathbf{a}_{U,i}$, $\mathbf{b}_{U,i}$, and $\mathbf{d}_{U,i}$ are the native Joint response, additive response, and defect projected into the carrier; the corresponding subscript Q denotes projection into the complement. $E_{add,U,i}$ and $E_{add,Q,i}$ are the two symmetric normalized composition errors.

any whitened response or defect vector, Eq. (I3) reports the squared-norm fraction in the carrier and complement and the corresponding enrichment relative to the 4/192 and 188/192 dimensional baselines.

$$\begin{aligned} f_U(\mathbf{v}) &= \frac{\| \mathbf{P}\mathbf{v} \|_2^2}{\| \mathbf{v} \|_2^2 + \varepsilon} \qquad f_Q(\mathbf{v}) = \frac{\| \mathbf{Q}\mathbf{v} \|_2^2}{\| \mathbf{v} \|_2^2 + \varepsilon} \\ enrich_U(\mathbf{v}) &= \frac{f_U(\mathbf{v})}{4/192} \qquad enrich_Q(\mathbf{v}) = \frac{f_Q(\mathbf{v})}{188/192} \end{aligned} \tag{I3}$$

Here, $\mathbf{v}$ is any whitened response or defect vector, $f_U(\mathbf{v})$ and $f_Q(\mathbf{v})$ are its squared-norm fractions in the carrier and complement, and $enrich_U$ and $enrich_Q$ divide those fractions by their dimensional baselines. The numerical floor is $\varepsilon = 10^{-12}$.

Enrichment is a dimension-normalized concentration diagnostic, not a significance test; high carrier enrichment does not imply that the carrier is the unique locus of composition.

I.3 Local route-conditioned recurrent operators

For each Factual, Single, and Joint route, consecutive factual-whitened hidden states from the 411 FIT units are stacked and fitted with an affine one-step least-squares operator:

$$\begin{aligned} \mathbf{Z}_r &= [\mathbf{X}_r, \mathbf{1}] \\ \mathbf{B}_r &= pinv\left(\mathbf{Z}_r; rcond = 10^{-12}\right) \mathbf{Y}_r \\ \begin{bmatrix} \mathbf{K}_r^\top \\ \eta_r^\top \end{bmatrix} &= \mathbf{B}_r \\ \widehat{\mathbf{Y}}_r &= \mathbf{X}_r \mathbf{K}_r^\top + \mathbf{1}\eta_r^\top \\ R_r^2 &= 1 - \frac{SSE_r}{SST_r} \end{aligned} \tag{I4}$$

Here, $r \in \{F, S, J\}$ is a route label rather than carrier rank; $\mathbf{X}_r$ and $\mathbf{Y}_r$ stack current and next whitened states, $\mathbf{Z}_r = [\mathbf{X}_r, \mathbf{1}]$ adds the intercept column, and $\mathbf{B}_r$ is the pseudoinverse least-squares solution with $rcond = 10^{-12}$. $\mathbf{K}_r$ is the fitted linear part and η_r its affine intercept. SSE_r, SST_r, and R_r^2 measure TEST fit quality.

The affine fit is evaluated on the 101 TEST units using the stated R-squared statistic and registered numerical floor.

Equation (I5) summarizes the Single-to-Joint operator difference and supporting normalized descriptors.

$$\begin{aligned} \Delta\mathbf{K}_{JS} &= \mathbf{K}_J - \mathbf{K}_S \\ d_{JS} &= \frac{\| \mathbf{K}_J - \mathbf{K}_S \|_F}{\| \mathbf{K}_F \|_F + \varepsilon} \\ \rho(\mathbf{K}) &= \max_i |\lambda_i(\mathbf{K})| \\ \nu(\mathbf{K}) &= \frac{\| \mathbf{K}^\top\mathbf{K} - \mathbf{K}\mathbf{K}^\top \|_F}{\| \mathbf{K} \|_F^2 + \varepsilon} \end{aligned} \tag{I5}$$

Here, $\Delta\mathbf{K}_{JS} = \mathbf{K}_J - \mathbf{K}_S$ is the fitted Joint-minus-Single operator difference, d_{JS} normalizes its Frobenius magnitude by the factual operator, $\rho(\mathbf{K})$ is spectral radius based on eigenvalues $\lambda_i(\mathbf{K})$, and $\nu(\mathbf{K})$ is a normalized non-normality diagnostic.

The matrix norm in Eq. (I5) is the Frobenius norm. These route-conditioned operators are local descriptive fits over the registered population, not exact recurrent dynamics, global Koopman operators, or reduced-state models.

I.4 CARRIER/COMPLEMENT BLOCKS, RETENTION, LEAKAGE, AND INWARD TRANSFER

For any fitted operator, Eq. (I6) separates carrier-to-carrier, carrier-to-complement, complement-to-carrier, and complement-to-complement blocks.

$$\begin{aligned} \mathbf{A}_{PP} &= \mathbf{PAP} \qquad \mathbf{A}_{PQ} = \mathbf{PAQ} \\ \mathbf{A}_{QP} &= \mathbf{QAP} \qquad \mathbf{A}_{QQ} = \mathbf{QAQ} \end{aligned} \tag{I6}$$

In Eq. (I6), the rightmost projector selects the input subspace and the leftmost projector selects the output subspace: $\mathbf{PAP}$ is carrier-to-carrier, $\mathbf{QAP}$ is carrier-to-complement, $\mathbf{PAQ}$ is complement-to-carrier, and $\mathbf{QAQ}$ is complement-to-complement.

For each route operator, Eq. (I7) defines carrier retention, outward leakage, and inward transfer.

$$\begin{aligned} R_{in} &= \frac{\| \mathbf{PKP} \|_F^2}{\| \mathbf{KP} \|_F^2 + \varepsilon} \\ L_{out} &= \frac{\| \mathbf{QKP} \|_F^2}{\| \mathbf{KP} \|_F^2 + \varepsilon} \\ G_{in} &= \frac{\| \mathbf{PKQ} \|_F^2}{\| \mathbf{KQ} \|_F^2 + \varepsilon} \end{aligned} \tag{I7}$$

Here, R_{in} is the fraction of carrier-originating fitted operator magnitude that returns to the carrier, L_{out} is the complementary fraction entering the orthogonal complement, and G_{in} is the fraction of complement-originating magnitude entering the carrier.

When the denominator is nonzero, retention plus outward leakage is one up to numerical precision. Low retention and high leakage mean that a carrier-originating one-step perturbation is dominated by coupling into the wider hidden state rather than remaining inside the rank-4 carrier.

I.5 DIMENSION-NORMALIZED OPERATOR BLOCKS

Because the four operator blocks have very different numbers of entries, Eq. (I8) reports block fractions, per-entry RMS magnitude, and dimension-normalized enrichment.

$$\begin{aligned} f_b &= \frac{\| \mathbf{A}_b \|_F^2}{\| \mathbf{A} \|_F^2 + \varepsilon} \\ RMS_b &= \frac{\| \mathbf{A}_b \|_F}{\sqrt{n_b}} \qquad RMS_{global} = \frac{\| \mathbf{A} \|_F}{192} \\ RMSenrich_b &= \frac{RMS_b}{RMS_{global} + \varepsilon} \\ EnergyEnrich_b &= \frac{f_b}{n_b / 192^2} \end{aligned} \tag{I8}$$

Here, $b \in \{PP, PQ, QP, QQ\}$ indexes an operator block, n_b is its number of entries, f_b is its squared-Frobenius fraction, RMS_b is per-entry RMS magnitude, and $RMSenrich_b$ and $EnergyEnrich_b$ compare block magnitude or squared-magnitude concentration with the full-operator baseline.

The carrier-to-carrier block contains 16 entries, each cross-subspace block contains 752 entries, and the complement-to-complement block contains 35344 entries. In EnergyEnrich, "energy" means squared Frobenius magnitude of the fitted operator block, not physical energy. These descriptors compare block concentration after accounting for dimension; they are not hypothesis tests.

I.6 FULL, BASELINE, AND REQUEST-ONLY JOINT PATCH GEOMETRY

For the rank-4 Single-only addressable map, Eq. (I9) decomposes the affine Joint coefficient into the full prediction, zero-edit baseline, and request-only component.

$$\begin{gathered}\mathbf{c}_{full} = f_{addr,4}\left(\mathbf{s}^F, \mathbf{e}_{xy}\right) \\ \mathbf{c}_{base} = f_{addr,4}\left(\mathbf{s}^F, \mathbf{0}\right) \\ \mathbf{c}_{request} = \mathbf{c}_{full} - \mathbf{c}_{base} \\ \delta\mathbf{z}_{full} = \mathbf{U}_4\mathbf{c}_{full} \qquad \delta\mathbf{z}_{base} = \mathbf{U}_4\mathbf{c}_{base} \qquad \delta\mathbf{z}_{request} = \mathbf{U}_4\mathbf{c}_{request}\end{gathered} \tag{I9}$$

Here, $\mathbf{c}_{full}$ is the direct affine coefficient used for rollout, $\mathbf{c}_{base}$ is the zero-edit affine baseline, and $\mathbf{c}_{request} = \mathbf{c}_{full} - \mathbf{c}_{base}$ isolates the request-dependent component. Multiplication by $\mathbf{U}_4$ gives the corresponding raw hidden-space patches.

The actual rollout always uses the full affine coefficient; the baseline and request-only terms are post-hoc geometry diagnostics only.

Accordingly, the full affine patch is the primary geometry object because it is the intervention actually used for rollout; the other components are secondary decompositions used only to interpret that patch.

Equation (I10) maps any raw-space perturbation into four-dimensional coordinates in the whitened carrier for comparison with the projected native Joint oracle from Eq. (D5a).

$$\begin{gathered}\mathbf{u}\left(\mathbf{v}\right) = \mathbf{Q}_U^\top \mathbf{W}_{wh}\mathbf{v} \\ E\left(\mathbf{v}, oracle\right) = \frac{2 \parallel \mathbf{u}\left(\mathbf{v}\right) - \mathbf{u}\left(\delta\mathbf{z}_{oracle}\right) \parallel_2}{\parallel \mathbf{u}\left(\mathbf{v}\right) \parallel_2 + \parallel \mathbf{u}\left(\delta\mathbf{z}_{oracle}\right) \parallel_2 + \varepsilon}\end{gathered} \tag{I10}$$

Here, $\mathbf{u}\left(\mathbf{v}\right) = \mathbf{Q}_U^\mathsf{T}\mathbf{W}_{wh}\mathbf{v} \in \mathbb{R}^4$ gives carrier coordinates of raw-space perturbation $\mathbf{v}$, $\delta\mathbf{z}_{oracle}$ is the projected native Joint oracle patch, and $E\left(\mathbf{v}, oracle\right)$ is their symmetric normalized coordinate distance.

Distances are computed on the same 101 TEST units. Distance to this particular projected native Joint oracle is a bounded geometry diagnostic and is not a sufficient proxy for rollout effectiveness.

I.7 INTERPRETATION BOUNDARY

These diagnostics support only bounded mechanistic interpretations: carrier enrichment does not make the rank-4 subspace the unique locus of composition; local operator coupling does not establish a globally invariant four-dimensional state; and oracle-distance results do not make hidden-space geometry irrelevant. All conclusions remain checkpoint-, population-, horizon-, and diagnostic-specific.