BEYOND ENDPOINT SUFFICIENCY: INNER-TIME TRAJECTORY AS A REPRESENTATION CHANNEL IN ITERATIVE SYSTEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Many iterative models treat the final converged state as the representation. We study endpoint insufficiency: convergence can erase task distinctions, and the trajectory can carry reliability information even when the endpoint remains the answer. We formalize representations as $\mathrm{Rep}(x) = \big(z_T(x), r(z_{0:T}(x))\big)$, give a contractive construction with coincident endpoints and a label-informative signed-area functional, and introduce audits with strong endpoint, residualization, permutation, and orientation controls. Exploratory KITTI reliability audits use frozen RAFT-Things and SEA-RAFT checkpoints not trained on KITTI. At 80% retained coverage, passive history reduces retained flow error by 8.1% beyond a learned endpoint-plus-convergence baseline on RAFT-Things; simple convergence reduces it by 4.8% beyond endpoint plus native uncertainty on SEA-RAFT. Richer geometry adds no resolved utility beyond convergence profiles, and the KITTI-finetuned RAFT checkpoint supplies a negative boundary. Controlled dynamics, looped Transformers, diffusion features, and a residual-null visual case expose distinct mechanisms and limits. These results support endpoint-sufficiency auditing, not universal trajectory access: retain history when it improves the declared target beyond strong inexpensive controls.

1 INTRODUCTION

Many iterative and implicit models produce representations through a convergent computation. Deep equilibrium models (DEQs) solve for a fixed point; recurrent networks update hidden states over time; iterative refinement models progressively transform an initial state into a final one. In most uses, the final state is treated as the representation and the path taken to obtain it is treated as an implementation detail.

We distinguish *answer insufficiency*, where convergence removes downstream-label distinctions, from *reliability insufficiency*, where history predicts whether the endpoint answer is trustworthy. Shared-map refinement offers a controlled answer-loss case: similar endpoints can follow different paths, a subtype we call endpoint collapse. Iterative predictors such as RAFT motivate the reliability case. These are tests of computation history, not claims about generic external-time memory.

We study the representation decomposition

$$z_{0:T}(x) \longmapsto \mathrm{Rep}(x) = \big(z_T(x),\ r(z_{0:T}(x))\big), \tag{1}$$

where z_T is the endpoint channel and $r(z_{0:T})$ summarizes how refinement settles. Our claim is narrow: endpoint-only representations can discard task distinctions or reliability information recoverable from computation history.

Making the phenomenon auditable. Final-output objectives need not reveal whether history carries complementary answer or reliability information. An endpoint-sufficiency audit fixes the model and target, strengthens endpoint access, and tests history against confidence, residualization, and matched perturbation controls. It must allow negative outcomes when the endpoint already suffices.

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.

Contributions. We give a sufficient witness for endpoint collision with path-label separation; define audits with strong endpoint, residualization, permutation, orientation, and reliability controls; and test selective reliability on RAFT-Things and SEA-RAFT beyond learned combined controls. Controlled chirality, scale-dependent WikiText gains, transductive SPair correspondence, and a residual-null visual case delimit distinct mechanisms. Failed richer-history and active-response tests are reported alongside positives. This is a diagnostic protocol, not a universally superior trajectory method.

2 Endpoint and Trajectory Channels

2.1 Contractive Iterative Representation

We consider shared-map recurrent refinement models of the form

$$z_{k+1} = F_\theta(z_k, x), \qquad k = 0, \dots, T-1, \tag{2}$$

where the same update map F_θ is reused at every inner refinement step for a fixed input x. In the regime we study, repeated application is contractive enough to yield a stable endpoint. In an endpoint-only model, a downstream readout uses z_T. In a trajectory-aware model, a downstream readout may also use a summary of the path:

$$\hat{y} = g\big(z_T,\ r(z_{0:T})\big). \tag{3}$$

In probe experiments, fixed low-dimensional r limits readout engineering and exposes interpretable feature groups; it is not claimed to be optimal. WikiText-103 separately tests learned fusion of already-computed loop states under a fixed rollout, with endpoint-repeat as the same-executed-head control. Plain endpoint bypasses that fusion head and is not active-capacity matched. For example, the path-functional probe uses endpoint state, mean inner-step displacement, path length, and final residual; the topology-conflict probe uses per-step pooled latent means and standard deviations and their differences. **Appendix A.3** gives exact readouts.

Contraction matters because it creates the tension tested here: repeated refinement makes endpoints stable and comparable, but the same compression can erase distinctions about how the endpoint was reached. Contraction is the clean model of endpoint collapse; SPair-71k and WikiText-103 test whether the endpoint–trajectory distinction remains measurable beyond that idealized regime.

Trajectory-axis convention. The clean theoretical object is the fixed-input inner-time trajectory $z_{0:T}$. Some standard-system audits use analogous computation histories rather than the exact contractive setting of Proposition 1: RAFT uses recurrent optical-flow update iterations, SPair uses diffusion timesteps, and WikiText uses loop states. The path-functional residual audit also conditions an external-position endpoint history $(z_{1,T}, \dots, z_{64,T})$. We name the audited axis explicitly in each setting because these histories are not interchangeable; the shared question is whether the chosen endpoint is sufficient for the target under test.

Proposition 1 (Geometric endpoint collapse with trajectory chirality)**.** *Let $0 < a < 1$, let $0 < \phi < \pi$, let $y \in \{-1, +1\}$, and let $c, v \in \mathbb{R}^2$ with $v \neq 0$. Define the planar fixed-input contractive system*

$$z_0 = c + v, \qquad z_{k+1} = c + aR_{y\phi}(z_k - c), \tag{4}$$

where $R_{y\phi}$ is the rotation matrix by angle $y\phi$. Then for every $k \geq 0$,

$$z_k(y) = c + a^k \big(R_{y\phi}\big)^k v, \tag{5}$$

so both classes share the same start $z_0 = c+v$ and converge to the same attractor c, with exponentially collapsing endpoint gap

$$\big\|z_K(+1) - z_K(-1)\big\| \leq 2a^K \|v\|. \tag{6}$$

Moreover, for every integer horizon $K \geq 2$, the chirality functional

$$\chi(z_{0:K}) = \sum_{k=0}^{K-2} \det\big(z_{k+1} - z_k,\ z_{k+2} - z_{k+1}\big) \tag{7}$$

satisfies $\operatorname{sign}\big(\chi(z_{0:K})\big) = y$.

Proof. **Appendix A.2** gives the full proof. The key step is that successive displacements $z_{k+1} - z_k$ are contracted rotations of one another by the same class-dependent angle $y\phi$, so the endpoint gap shrinks as a^K while each signed-area term in χ carries the class-dependent sign y. Thus the endpoint margin collapses exponentially while a genuinely path-dependent functional (a non-telescoping signed area of consecutive displacement pairs) remains label-informative. This stylized geometric setting shows that terminal-state sufficiency can fail under shared contraction even when the trajectory channel is not a simple initial–final difference. The appendix also gives a local nonlinear sufficient condition: for C^2 contractive maps with a common attracting fixed point and an oriented two-dimensional mode in the linearization, the same signed-area separation persists in a sufficiently small neighborhood of the attractor. □

Shrinking endpoint distance alone does not imply finite-horizon information loss: distinct exact endpoints may remain separable. For any chosen $K \geq 2$, setting $\phi = \pi/K$ gives an exact collision, $z_K(+1) = z_K(-1) = c - a^K v$, while χ retains opposite signs. For generic angles, the bound concerns endpoint margin, not impossibility of exact decoding.

The propositions are sufficient witnesses, not models of arbitrary learned dynamics. Different attractors leave non-decaying separation, and nonlinear chirality requires its quadratic margin to dominate the cubic remainder (**Appendix A.2**). A planar mode can embed in higher dimensions. Audits extend to untied Transformer depth, RNNs, Neural ODEs (Chen et al., 2018), and DEQs, but their dynamics need not satisfy these assumptions.

3 Audit Protocol: Testing Endpoint Sufficiency

An endpoint-sufficiency audit asks whether z_T alone exposes the target-relevant information. We use "signal" in this operational sense: label-predictive information for answer-insufficiency audits and failure-predictive information for reliability-insufficiency audits. We use three experimental roles: answer-insufficiency audits ask whether history recovers answer signal lost by the endpoint; reliability-insufficiency audits ask whether history predicts endpoint failure beyond confidence baselines; and mechanism checks identify the audited trajectory-axis and which kind of trajectory information survives or fails.

Answer-insufficiency audits. These endpoint-collapse probes use inner-time refinement, external labels such as controlled latent factors or transient cues, and test regimes where endpoints overlap, collapse, or become nuisance-dominated. The controls distinguish trajectory information from cheap alternatives: endpoint-strength probes, shuffled or permutation-augmented trajectories, shortcut readouts, and endpoint-residualized audits when raw states are available.

Reliability-insufficiency audits. Here the endpoint is the task answer, and the target is whether that answer is wrong or unreliable. The competing explanation is endpoint confidence rather than answer-label recovery, so the controls are endpoint-only confidence features, endpoint-overmatched learned endpoint-only baselines, last-update controls, simple convergence-variance baselines inspired by Durasov et al., 2024, and richer history probes. Report each metric and comparator separately: improvement over an endpoint probe need not imply improvement over a convergence baseline on every metric. Dense-output intervals must preserve physical frame pairs across repeated appearances, not assume independent pixels or split-specific image IDs. Chronology claims additionally require controls that preserve cross-stream alignment and change order alone; the RAFT and SPair shuffles do not isolate it.

Branch-specific endpoint controls. For answer-insufficiency probes, endpoint-strength controls use fixed, label-blind expansions of z_T to match feature count without adding learned endpoint decoders. For reliability audits, where the target is endpoint failure and endpoint confidence is the main competing explanation, we instead allow endpoint-overmatched learned endpoint-only baselines, including calibrated MLPs, but still forbid access to intermediate states or update-history features. Thus endpoint-overmatch is branch-specific: fixed endpoint expansions for answer audits, learned endpoint-only confidence baselines for reliability audits.

Operational guidance. Test trajectory access when contraction may remove answer-relevant distinctions, or when settling behavior may predict failure beyond endpoint confidence. Use strong endpoint controls and keep history only if it improves the audited target; increased endpoint capacity can remove the gain. Residualization, permutation, and orientation controls then delimit the supported mechanism, not a universal requirement to expose trajectories.

The post-hoc audits freeze a base model or cached trajectory source before fitting readouts. WikiText-103 instead trains each backbone/readout configuration end-to-end with the same loop count and within-setting training budget; it does not isolate readout access on identical learned states. In answer-insufficiency probes, the readout family is endpoint-only z_T, endpoint-strength $\psi(z_T)$, trajectory $(z_T, r(z_{0:T}))$, and shuffled trajectory $(z_T, r(z'_{0:T}))$ from another example in the same split. Endpoint-strength controls are endpoint-only feature-map controls, not learned hidden networks over z_T: in dimension-matched variants, fixed label-blind $\psi(z_T) \in \mathbb{R}^D$ has the same feature dimension D as the trajectory feature vector, so the fitted linear classifier has the same input dimension. These controls test feature-count and endpoint-readout-strength loopholes while preserving the endpoint-only information source; **Appendix A.1** maps all audit roles. **Appendix A.3** lists controlled-probe readouts, and **Appendix A.12** gives RAFT feature definitions. Unless otherwise stated, reported stochastic metrics are mean $\pm$ standard deviation over repeated probe seeds.

Answer audits use full test sets and diagnostic subsets ranked by collision severity, stable drive, or endpoint proximity. The path-functional "endpoint bands" are an exception: all raw endpoints are zero, so these are tie-selected subsets, not a hardness ladder (**Appendix A.5**). The primary metric is AUROC, with particular attention to

$$\Delta_{\text{traj}} = \text{AUROC}(P_{\text{traj}}) - \text{AUROC}(P_{\text{endpoint}}). \tag{8}$$

When a table reports a paired classification-accuracy gap, that quantity is computed from thresholded correctness on the same test examples and is used as a descriptive paired diagnostic; it is not an AUROC difference.

4 Practical Reliability: Flow Failure Detection

Does history detect more flow failures at a fixed rejection budget? We freeze RAFT-Things (Teed & Deng, 2020) at 24 updates and SEA-RAFT-M (Wang et al., 2024) at four refinements, neither KITTI-trained, and retain KITTI-finetuned RAFT as a continuity control. The target is the KITTI-2015 outlier criterion (Menze & Geiger, 2015).

All use one 120/40/40 fit/validation/test pair split and 8,192 uniformly sampled valid test pixels per pair. Validation average precision (AP) selects among two linear and two nonlinear candidates per readout. Observed endpoint E includes flow, spatial features, and final internal norms, not the full hidden tensor. Baseline B adds last-update, variance, and update-energy summaries; SEA-RAFT also includes native uncertainty U. History H adds convergence profiles, temporal Gram, and signed area. Thus the comparator combines endpoint and simple convergence, rather than testing them separately.

Table 1: **Selective reliability on 40 KITTI pairs at 80% retained coverage.** Higher AP/recall and lower EPE are better. B: combined baseline; H: passive history; $E + U$: endpoint plus native uncertainty. Appendix A.13 reports intervals and failed primary tests.

Checkpoint	Readout	AP	EPE80 (px)	Failure recall
RAFT-Things	B	0.8157	1.0004	83.69%
RAFT-Things	$B + H$	0.8345	0.9196	85.66%
SEA-RAFT	$E + U$	0.7832	0.8854	85.33%
SEA-RAFT	B	0.7920	0.8429	86.79%
SEA-RAFT	$B + H$	0.7941	0.8361	86.80%
RAFT-KITTI	B	0.3005	0.2379	92.23%
RAFT-KITTI	$B + H$	0.3098	0.2349	92.39%

On RAFT-Things, adding H improves AP by 0.0188 (95% CI $[0.0092, 0.0326]$) and reduces retained EPE by 0.0808 pixels ($[-0.2262, -0.0125]$), or 8.1%. On SEA-RAFT, adding simple convergence

to $E + U$ improves AP by 0.0088 ($[0.0029, 0.0146]$) and reduces retained EPE by 0.0425 pixels ($[-0.0890, -0.0046]$), or 4.8%.

History beyond B is unresolved on SEA-RAFT (AP $+0.0022$, CI $[-0.0005, 0.0054]$) and RAFT-KITTI ($+0.0093$, $[-0.0148, 0.0325]$). Richer geometry adds no resolved benefit beyond convergence profiles. These positives were prespecified secondary tests; the primary RAFT active-response and SEA-RAFT richer-history tests remain negative. No chronology or necessary geometric mechanism is inferred.

Intervals use 600 paired physical-frame draws, conditional on this fit and unadjusted for multiplicity. Previously inspected KITTI data provide exploratory, not scene-disjoint or cross-family, evidence. Global pixel rejection improves retained reliability, not full-coverage flow or certified deployment efficiency. **Appendix A.13** gives all readouts, prevalence, and verification. The corrected legacy comparison and normalization disclosure remain in **Appendix A.12**; new caches use official forward.

5 Controlled Calibration: Answer Insufficiency

This section calibrates the answer-insufficiency branch of the audit. The endpoint-collapse ladder shows that refinement-history signal can survive when endpoints collide; the path-functional probe gives the cleanest endpoint-sufficiency stress test; and the final scope check asks whether the separation is tied to a single update mechanism.

5.1 Endpoint-Collapse Calibration

The controlled endpoint-collision construction is an existence proof for endpoint insufficiency. It uses a scalar contraction with a label-independent nuisance fixed point:

$$y_i \sim \text{Bernoulli}(1/2), \qquad c_i \sim \text{Uniform}[-3, 3], \qquad a_i = a_{y_i}, \quad a_0 = 0.2, \quad a_1 = 0.9,$$

$$z_{i,0} = c_i + 1, \qquad z_{i,k+1} = a_i z_{i,k} + (1 - a_i)c_i, \quad k = 0, \ldots, 39.$$

Thus $z_{i,K} = c_i + a_i^K$ for $K = 40$: both labels converge toward endpoints dominated by c_i, while contraction rate remains label-dependent. This separates endpoint content from settling dynamics in the simplest possible setting. Endpoint-only AUROC is 0.5024 ± 0.0102, while the fixed settling-history summary $[z_T, \mu(\Delta z), \text{var}(\Delta z), \sum_k |\Delta z_k|]$ reaches 1.0000 ± 0.0000; an endpoint-dimension control remains at chance. Since this toy uses aggregate statistics rather than ordered raw trajectories, we treat it as a possibility result and use stricter matched-trajectory controls below.

We then use learned versions of the toy to separate the source of endpoint overlap from the readout test. In the learned-collision run, a shared-map model is trained on a collision-rich synthetic classification task: a stable cue makes the endpoint task learnable, while a transient cue creates opposite-label examples that can land near one another in endpoint space. This run asks whether endpoint overlap appears after training, rather than being imposed analytically.

The stricter frozen-backbone run then freezes the trained update map and fits only post-hoc readouts on cached states. Here endpoint, trajectory, shuffled-trajectory, and endpoint-expanded probes all see the same fixed dynamics; only the readout input changes. For a test example i, we standardize final endpoints over the test split and define endpoint-collision severity as

$$s_i = \min_{j:\, y_j \neq y_i} \|\tilde{z}_{T,i} - \tilde{z}_{T,j}\|_2,$$

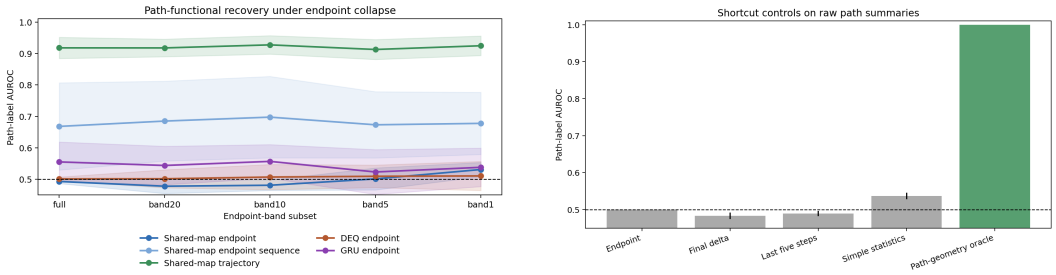
so smaller s_i means a closer opposite-label endpoint neighbor and therefore a more severe collision. This turns endpoint collapse from a constructed toy into a learned frozen-representation question: as final states become locally confusable, does the refinement history degrade more slowly? On the hardest 1% learned-collision slice, the endpoint probe reaches 0.8170 ± 0.0475 AUROC while the refinement-history summary reaches 0.8990 ± 0.0304. In the stricter frozen-backbone study, Hard-1% endpoint reaches 0.8217, the refinement-history summary reaches 0.8754, and the shuffled-history control falls to 0.8159; a stronger endpoint-expanded control using only z_T also does not close the gap. Thus endpoint collapse can emerge after training, and the gain depends on matched per-example refinement-history information rather than feature count alone. Section 6 returns to this ladder under endpoint residualization.

An input-modulation check keeps both variants time-tied and endpoint task performance similar. The unmodulated map has larger Δ_{traj}: 0.0239 and 0.0552 on nominal Hard 5% and 1%, versus 0.0160 and 0.0293. The latter slice is floored at 50 of 2,500 cases. Shared-minus-modulated gaps are 0.00795 (95% seed-bootstrap CI $[0.0042, 0.0119]$) and 0.0259 ($[0.0071, 0.0551]$). Added parameters and variant-specific slices prevent isolating shared-weight causality; this is supportive input-conditioning evidence (**Appendix A.4**).

5.2 Path-Functional Answer Insufficiency

Path-functional collapse probe. The path-functional probe deliberately separates backbone training from the audited label. Each example has a stable nuisance label s_i for training and an independent path label y_i for frozen probes. The path channel distinguishes long sign runs from oscillation; centering/scaling fixes raw endpoints to zero and weakens tested shortcut cues. This is not a memory task trained on y_i. **Appendix A.5** specifies generation, input access, and controls.

Full-set endpoint AUROCs are 0.4932 ± 0.0065 (shared map) and 0.5553 ± 0.0634 (GRU). DEQ scores 0.5013 ± 0.0060 but receives only the stable channel: it is a no-path-access control, not evidence of erasure. The endpoint sequence $(z_{1,T}, \ldots, z_{64,T})$ reaches 0.6682 ± 0.1386, versus trajectory 0.9180 ± 0.0337 and shuffled 0.4981 ± 0.0049. The trajectory-minus-sequence gap is 0.2498; the saved range $[0.1099, 0.3455]$ is a descriptive seed-quantile range, not a confidence interval. Raw shortcuts remain weak (final delta 0.4833, last-five window 0.4890, simple statistics 0.5367). Section 6 conditions endpoint history on the final state.



(a) Frozen readouts

(b) Leak controls

Figure 1: **Path-functional frozen probes.** Trajectory outperforms the tested endpoint and shortcut probes. Plotted endpoint bands are tie-selected subsets, not a hardness ladder; DEQ has no path input.

The same-backbone separation persists at update budgets $K \in \{2, 4, 8, 16\}$: trajectory AUROC ranges from 0.9696 ± 0.0049 at $K = 2$ to 0.7881 ± 0.0729 at $K = 16$. Endpoint and history use the same rollout; equal update counts across different cells do not equalize FLOPs or readout cost (**Appendix A.8**).

5.3 Scope Across Update Mechanisms

The GRU, DEQ, and linear SSM runs are scope controls for the answer-insufficiency branch, not benchmark comparisons: they test whether endpoint/trajectory separation depends on the shared-map recurrent cell or also appears under related recurrent, equilibrium, and state-space update mechanisms. The cross-model transient-signal washout stress test asks whether the endpoint–trajectory separation survives when multiple backbones solve the same stable task. Backbones train only on a persistent stable cue; frozen probes are then evaluated on a transient cue injected at the first inner step and washed out by stronger stable drives (**Appendix A.9**). On the top 1% washout slice, endpoint-only probes approach chance for shared-map (0.4863 ± 0.0883), DEQ (0.5147 ± 0.0634), and GRU (0.5317 ± 0.0692), while the shared-map trajectory reaches 0.8953 ± 0.1104. GRU trajectory access reaches 0.6817 ± 0.0981. As an additional backbone check, a linear SSM under the path-functional protocol gives endpoint 0.587 ± 0.030, trajectory 0.9995 ± 0.0006, and shuffled 0.575 ± 0.045 (**Appendix A.10**). These controls show the diagnostic is not confined to a single update mechanism.

6 RESIDUAL AUDIT AND BOUNDARY CASES

Raw trajectory readouts can beat endpoint readouts because the path contains history missing from the endpoint, or because it gives a probe-friendly encoding of endpoint-recoverable information. We investigate these possibilities with an endpoint-conditioned residual audit. The controlled and WikiText residual audits use train-only affine fits; SPair instead fits unlabeled evaluation features and is explicitly transductive. For each earlier state, fit a ridge map from the final endpoint and subtract the endpoint-predictable component, $\tilde{z}_t = z_t - (W_t z_T + b_t)$. A *residual-positive* result means the tested probe recovers signal after affine endpoint conditioning. A *residual-null* result is consistent with endpoint-recoverability, but can also reflect probe or fitting limitations: it does not prove endpoint sufficiency. These terms apply to residualized answer-insufficiency audits; reliability audits follow Section 3, comparing each reliability metric against endpoint confidence, last-update, and convergence-variance controls separately.

The settings provide residual-positive controls, a chirality check, and a residual-null boundary, without identifying information-theoretic sufficiency.

Table 2: **Residual-audit reader map.** Outcomes are relative to the specified affine conditioning and finite probes.

Setting	Audit outcome	Interpretation
Path-functional	ordered residual $>$ permutation	residual-positive answer history
Learned collision	ordered residual $>$ permutation	residual-positive inner-time history
Orientation audit	wedge/Lévy $>$ permuted/reversed	time-antisymmetric chirality signal
Topology-conflict	ordered residual $\approx$ permutation	residual-null boundary

Ridge removes only affinely endpoint-predictable content; finite probes cannot rule out all nonlinear endpoint decoders. Bracketed accuracy-gap bounds below average within-seed bootstrap endpoints. They are descriptive, not confidence intervals for the across-seed mean (**Appendix A.6**).

For true fixed-input refinement, the residualizer is $\hat{z}_t = W_t z_T + b_t$ for inner-time steps $t < T$; for the path-functional residual audit, it conditions each external-position endpoint $z_{j,T}$ on $z_{64,T}$. The *ordered residual readout* concatenates residual states in their original order. The controlled *permutation-augmented control* resamples a common random permutation per training batch, randomizing each example's ordering without a fixed ordered code.

Residual-positive controls. The controlled probes are positive cases: ordered residual history survives endpoint conditioning. On the path-functional probe, the final external-position endpoint linearly predicts most of the ordered endpoint history ($R^2(z_{j,T} \mid z_{64,T}) \in [0.995, 0.998]$), yet a nonlinear endpoint MLP with two 128-unit ReLU hidden layers remains at 0.499 ± 0.012 AUROC. The ordered endpoint-history residual readout reaches 0.863 ± 0.097, while permutation-augmented residuals reach 0.509 ± 0.005; paired classification-accuracy gaps over endpoint MLP and permutation control are 0.295 $[0.271, 0.320]$ and 0.302 $[0.278, 0.326]$. **Appendix A.6** gives the full controls; finite nonlinear probes do not rule out every endpoint decoder.

On the learned collision ladder, where the audited states are true inner-time refinement states, ordered residual reaches 0.998 ± 0.001 full-set AUROC versus 0.841 ± 0.051 for permutation-augmented residuals, with paired classification-accuracy gap 0.207 $[0.191, 0.223]$. A label-permutation sanity check returns chance. **Appendix A.6** reports full residualized tables and hard slices.

Mechanism-specific positive: time-antisymmetric signal. Proposition 1 uses signed area as a non-telescoping trajectory witness. We test matching high-dimensional orientation features on the path-functional probe: a local-wedge feature that sums consecutive update areas, and a Lévy-area feature that sums pairwise update areas (Bonnier et al., 2019; Kidger & Lyons, 2021). Endpoint AUROC is 0.4932 ± 0.0065, while local wedge reaches 0.9733 ± 0.0018 and Lévy area reaches 0.9729 ± 0.0015. Step-permuted controls collapse to 0.5113 ± 0.0073 and 0.5077 ± 0.0026; after ridge endpoint residualization, local wedge remains 0.9214 ± 0.0110 while residual step-permutation

is 0.5040 ± 0.0074. Testing a forward-trained wedge probe on time-reversed trajectories drops to 0.1426 ± 0.1652, consistent with the antisymmetric sign flip; **Appendix A.7** gives full controls.

Residual-null boundary. The topology-conflict visual probe has a raw trajectory gain, but ordered residual AUROC is 0.484 ± 0.008 versus 0.485 ± 0.019 for the permutation-augmented control. One explanation is that intermediate states offer easier views of endpoint-predictable structure; the high ridge predictivity is consistent with this. The tested residual probes recover no advantage, but this does not establish that all useful history is endpoint-recoverable. Nor must a genuine history gain require order: unordered summaries can carry additional signal.

7 Mechanism-Dependent Standard Checks

SPair-71k: diffusion residual scope check. SPair-71k (Min et al., 2019) is a transductive correspondence check, not a new descriptor: 11-timestep concatenation and the best single timestep give nearly identical raw $\mathrm{PCK}_{\mathrm{bbox}}$ (0.6356 versus 0.6354). The reference timestep ($t = 2$) is selected on evaluation performance and is an oracle baseline. Ridge residualization and orientation PCA fit unlabeled evaluation-image features, not a disjoint training split. Residual concatenation reaches 0.1865, versus 0.1434 for mean pooling and 0.1117 for independent-image shuffling. A shared timestep permutation preserves descriptor dot products; independent shuffles disrupt cross-image alignment. Thus this comparison supports aligned residual correspondence, not isolated chronology. The strongest timestep appears to capture most of the raw task utility; signed-area features do not improve the residual descriptor. **Appendix A.14** reports the transductive protocol and dependence limitations.

WikiText-103: loop-state scale check. WikiText-103 trains endpoint-only, endpoint-repeat, and trajectory configurations end-to-end, each with six loops (Merity et al., 2017). Endpoint-repeat executes the same fusion head as trajectory; plain endpoint bypasses it, leaving 221,760 allocated parameters inactive at $d = 192$. Small-model test perplexities are 135.97 ± 1.51, 138.44 ± 0.80, and 127.56 ± 1.46, respectively, across 10 seeds. The gain therefore remains against the executed-head control, not only the smaller active endpoint. Orientation probes on trained loop states also beat step-permuted controls. At $d = 384$, endpoint and trajectory are close (82.23 ± 1.28 versus 82.33 ± 0.05), while endpoint-repeat is worse (86.87 ± 0.70). The larger run also doubles training steps, so it is a width/training-budget boundary, not an isolated capacity intervention or proof of sufficiency. **Appendix A.15** separates allocated and active capacity.

8 Related Work

Universal Transformers reuse computation over depth (Dehghani et al., 2019); DEQs explicitly solve for an endpoint (Bai et al., 2019; Winston & Kolter, 2020); and Path Independent Equilibrium Models connect initialization-independent settling with generalization (Anil et al., 2022). Our endpoint audit complements hidden-layer probes (Alain & Bengio, 2016; Belinkov et al., 2017; Tenney et al., 2019), tests of learned syntactic competence (Linzen et al., 2016), looped-transformer computation (Saunshi et al., 2025), and diffusion correspondence (Luo et al., 2023; Tang et al., 2023). None of these citations establishes our empirical endpoint-insufficiency claim by itself.

Reliability precedents include convergence-based uncertainty (Durasov et al., 2024) and flow/correspondence uncertainty (Wannenwetsch et al., 2017; Ilg et al., 2018; Truong et al., 2021); SciFlow instead improves iterative flow refinement (Lin et al., 2024). We test history beyond combined endpoint/convergence controls and SEA-RAFT uncertainty, not invent convergence-based uncertainty. Signature methods encode streams (Bonnier et al., 2019; Kidger & Lyons, 2021); our wedge and Lévy-area features are diagnostic special cases, not a new signature transform. The audited axis is inference-time computation, distinct from NTK training dynamics (Jacot et al., 2018) and external-time state-tracking expressiveness (Merrill et al., 2024).

9 Discussion and Limitations

The endpoint-plus-trajectory decomposition is a diagnostic framework, not a single-mechanism account. Controlled dynamics support chirality; corrected RAFT/SEA-RAFT support checkpoint-dependent convergence reliability; WikiText exposes a negative scale boundary; SPair supports transductive timestep-aligned residual correspondence without a useful signed-area extension; and the visual boundary is residual-null. RAFT aggregates are not strictly time-symmetric, and the RAFT/SPair shuffles do not isolate chronology. These distinctions constrain interpretation rather than turn every positive comparison into the same mechanism (**Appendix A.1**).

The strongest mechanism evidence is controlled. The practical audits use three fixed checkpoints in the RAFT family, one common pair split, and a previously inspected dataset. Two checkpoints were not KITTI-trained; the KITTI-finetuned control was. Scene disjointness is not established. GMA (Jiang et al., 2021) and FlowFormer (Huang et al., 2022) remain future checks. Feature/readout overhead is not an end-to-end deployment benchmark, and secondary positives do not reverse failed primary tests. The local theorem's common-attractor and oriented-mode assumptions can fail; SPair needs disjoint fitting before an inductive claim, and its reused images limit pair-row bootstrap interpretation. Finite probes do not certify information-theoretic insufficiency or safety-calibrated uncertainty. Some fixed statistics stream with $O(d)$ extra state, but full-history readouts, shuffling, and trajectory training can require $O(Td)$ storage and weaken endpoint-only memory advantages. Appendix A.12 distinguishes verified corrected statistics from incomplete remote-cache provenance. Positive and negative outcomes should both be reported.

10 Conclusion

Endpoint insufficiency names a representational risk: stable final states can erase path-visible distinctions or hide reliability information. Our controlled and standard-system audits test this risk while exposing readout limitations and mechanism-specific nulls. Practical gains can come from simple convergence rather than richer geometry. Endpoint sufficiency is an empirical assumption to test, not a default to inherit.

AI USE STATEMENT

Generative AI tools, including ChatGPT/Codex, Claude, and Gemini, assisted with conceptual framing, hypothesis and experiment design, mathematical reasoning and proof drafting, synthetic-data and experiment code, debugging, interpretation of results, literature exploration, and manuscript drafting and revision. AI-assisted numerical claims were checked against saved experimental artifacts where available; the scope and remaining verification limits are stated in Appendices A.12 and A.13. The authors take responsibility for the final text, claims, code, and results, including material developed with AI assistance.

ETHICS STATEMENT

This work uses synthetic data and public machine-learning benchmarks; it does not involve human-subject experiments. Endpoint-sufficiency audits may improve reliability assessment, but post-hoc scores can also create false confidence. The experiments do not validate safety-critical deployment or calibrated failure probabilities. Third-party datasets and checkpoints remain subject to their original licenses and access conditions, as discussed in Appendix A.3.

REPRODUCIBILITY STATEMENT

Appendix A.2 gives the theoretical assumptions and proofs. Appendix A.3 specifies readouts, resources, and third-party assets; subsequent appendices describe the controlled and standard-system protocols. Appendix A.12 records the corrected RAFT preprocessing, evaluation population, physical-pair bootstrap, metric definitions, and remaining cache-provenance limitations. SPair's transductive fitting and WikiText's allocated-versus-active capacity distinction are retained explicitly rather than presented as independently resolved controls. Appendix A.13 records the additional fixed-checkpoint audits, official-forward checks, matched readout selection, and complete saved-score reconstruction of the selective metrics and 600 physical-pair draws. Their positive secondary and negative primary outcomes are both reported. These are exploratory studies, not untouched-data confirmation.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding Intermediate Layers Using Linear Classifier Probes. *arXiv preprint arXiv:1610.01644*, 2016. URL https://arxiv.org/abs/1610.01644.

Cem Anil, Ashwini Pokle, Kaiqu Liang, Johannes Treutlein, Yuhuai Wu, Shaojie Bai, J. Zico Kolter, and Roger Grosse. Path Independent Equilibrium Models Can Better Exploit Test-Time Computation. In *Advances in Neural Information Processing Systems*, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/331c41353b053683e17f7c88a797701d-Abstract-Conference.html.

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. Deep Equilibrium Models. In *Advances in Neural Information Processing Systems*, 2019. URL https://arxiv.org/abs/1909.01377.

Yonatan Belinkov, Lluís Màrquez, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. Evaluating Layers of Representation in Neural Machine Translation on Part-of-Speech and Semantic Tagging Tasks. In *International Joint Conference on Natural Language Processing*, 2017. URL https://aclanthology.org/I17-1001/.

Patric Bonnier, Patrick Kidger, Imanol Perez Arribas, Cristopher Salvi, and Terry Lyons. Deep Signature Transforms. In *Advances in Neural Information Processing Systems*, 2019. URL https://arxiv.org/abs/1905.08494.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural Ordinary Differential Equations. In *Advances in Neural Information Processing Systems*, 2018. URL https://arxiv.org/abs/1806.07366.

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Łukasz Kaiser. Universal Transformers. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1807.03819.

Nikita Durasov, Doruk Oner, Jonathan Donier, Hieu Le, and Pascal Fua. Enabling Uncertainty Estimation in Iterative Neural Networks. In *International Conference on Machine Learning*, 2024. URL https://proceedings.mlr.press/v235/durasov24a.html.

Zhaoyang Huang, Xiaoyu Shi, Chao Zhang, Qiang Wang, Ka Chun Cheung, Hongwei Qin, Jifeng Dai, and Hongsheng Li. FlowFormer: A Transformer Architecture for Optical Flow. In *European Conference on Computer Vision*, 2022. URL https://doi.org/10.1007/978-3-031-19790-1_40.

Eddy Ilg, Özgün Çiçek, Silvio Galesso, Aaron Klein, Osama Makansi, Frank Hutter, and Thomas Brox. Uncertainty Estimates and Multi-Hypotheses Networks for Optical Flow. In *European Conference on Computer Vision*, 2018. URL https://arxiv.org/abs/1802.07095.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. In *Advances in Neural Information Processing Systems*, 2018. URL https://arxiv.org/abs/1806.07572.

Shihao Jiang, Dylan Campbell, Yao Lu, Hongdong Li, and Richard Hartley. Learning To Estimate Hidden Motions With Global Motion Aggregation. In *IEEE/CVF International Conference on Computer Vision*, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Jiang_Learning_To_Estimate_Hidden_Motions_With_Global_Motion_Aggregation_ICCV_2021_paper.html.

Patrick Kidger and Terry Lyons. Signatory: Differentiable Computations of the Signature and Logsignature Transforms, on Both CPU and GPU. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2001.00706.

Jamie Menjay Lin, Jisoo Jeong, Hong Cai, Risheek Garrepalli, Kai Wang, and Fatih Porikli. SciFlow: Empowering Lightweight Optical Flow Models with Self-Cleaning Iterations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2024. URL https://arxiv.org/abs/2404.08135.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies. *Transactions of the Association for Computational Linguistics*, 4: 521–535, 2016. URL https://aclanthology.org/Q16-1037/.

Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion Hyperfeatures: Searching Through Time and Space for Semantic Correspondence. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2305.14334.

Moritz Menze and Andreas Geiger. Object Scene Flow for Autonomous Vehicles. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. URL https://openaccess.thecvf.com/content_cvpr_2015/html/Menze_Object_Scene_Flow_2015_CVPR_paper.html.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer Sentinel Mixture Models. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1609.07843.

William Merrill, Jackson Petty, and Ashish Sabharwal. The Illusion of State in State-Space Models. In *International Conference on Machine Learning*, 2024. URL https://proceedings.mlr.press/v235/merrill24a.html.

Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. SPair-71k: A Large-scale Benchmark for Semantic Correspondence. *arXiv preprint arXiv:1908.10543*, 2019. URL https://arxiv.org/abs/1908.10543.

Nikunj Saunshi, Nishanth Dikkala, Zhiyuan Li, Sanjiv Kumar, and Sashank J. Reddi. Reasoning with Latent Thoughts: On the Power of Looped Transformers. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2502.17416.

Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent Correspondence from Image Diffusion. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2306.03881.

Zachary Teed and Jia Deng. RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. In *European Conference on Computer Vision*, 2020. URL https://arxiv.org/abs/2003.12039.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT Rediscovers the Classical NLP Pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4593–4601, 2019. URL https://aclanthology.org/P19-1452/.

Prune Truong, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning Accurate Dense Correspondences and When to Trust Them. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. URL https://arxiv.org/abs/2101.01710.

Yihan Wang, Lahav Lipson, and Jia Deng. SEA-RAFT: Simple, Efficient, Accurate RAFT for Optical Flow. In *European Conference on Computer Vision*, 2024. URL https://arxiv.org/abs/2405.14793.

Anne S. Wannenwetsch, Margret Keuper, and Stefan Roth. ProbFlow: Joint Optical Flow and Uncertainty Estimation. In *IEEE International Conference on Computer Vision*, 2017. URL https://openaccess.thecvf.com/content_iccv_2017/html/Wannenwetsch_ProbFlow_Joint_Optical_ICCV_2017_paper.html.

Ezra Winston and J. Zico Kolter. Monotone Operator Equilibrium Networks. In *Advances in Neural Information Processing Systems*, 2020. URL https://arxiv.org/abs/2006.08591.

A Supplementary Experiment Details

A.1 Audit Map and Control Roles

Table 3: **Audit map and control questions.** Each component targets a specific endpoint-sufficiency control question. The framework is unified by the endpoint-sufficiency question, not by a single path geometry. A complementary time-reversal view separates odd/time-antisymmetric modes (the theorem, wedge/Lévy-area, and WikiText orientation audit), aggregate convergence reliability (RAFT), transductive aligned residual correspondence (SPair), and a residual-null boundary (topology-conflict).

Component	Purpose	Trajectory-axis	Audit branch / role
Proposition 1	formal witness	inner-time: fixed-input iterates	answer: formal witness
Local nonlinear proposition	robustness witness	inner-time: local iterates	answer: local witness
Controlled scalar collision	possibility calibration	inner-time: scalar contraction	answer: calibration
Learned / frozen collision	learned calibration	inner-time: shared-map states	answer: learned calibration
Shared-dynamics ablation	mechanism check	inner-time: shared/modulated updates	answer: mechanism check
Path-functional probe	stress test	path-position: inner summaries	answer: stress test
Cross-model washout, SSM, GRU	backbone scope	inner-time: update mechanisms	answer: backbone scope
Path-functional residual audit	controlled endpoint-recoverability test	external-position: endpoint history	answer: residualized control
Learned-collision residual audit	learned-state endpoint-recoverability test	inner-time: learned states	answer: residualized control
Orientation audit	mechanism-specific positive	inner-time: displacement sequence	mechanism: orientation
Topology-conflict	negative boundary	inner-time: reconstruction iterates	boundary: residual-null
RAFT/KITTI	utility anchor	inner-time: RAFT flow updates	reliability: utility audit
SEA-RAFT/KITTI	native-uncertainty control	inner-time: flow refinements	reliability: positive simple/negative richer history
SPair-71k	standard-system scope	diffusion-time: saved timesteps	transductive residual scope
WikiText-103	readout and scale check	loop-time: Transformer states	standard-system: readout/scale scope

Table 4: **Path signals and control questions.** Companion to Table 3; all components and qualifications are retained.

Component	Path signal tested / found	Control question
Proposition 1	odd / time-antisymmetric signed area	rules out convergence implying endpoint sufficiency
Local nonlinear proposition	odd / time-antisymmetric projected chirality	rules out a purely linear-rotation witness
Controlled scalar collision	even / time-symmetric settling rate	isolates nuisance-attractor endpoint domination
Learned / frozen collision	low-dimensional refinement-history summary	rules out construction-only endpoint overlap
Shared-dynamics ablation	input-modulation sensitivity	supportive, not isolated shared-weight causality
Path-functional probe	global path-shape history	rules out endpoint-sequence and shortcut controls
Cross-model washout, SSM, GRU	transient washout history	rules out a single-cell diagnostic artifact
Path-functional residual audit	ordered residual history	tests signal after affine endpoint conditioning
Learned-collision residual audit	ordered residual history	tests learned histories after endpoint conditioning
Orientation audit	odd / time-antisymmetric wedge and Lévy-area	tests chirality as an empirical mechanism
Topology-conflict	raw gain; residual-null	tests whether the specified residual probe retains signal
RAFT/KITTI	aggregate history; not strictly time-symmetric	AUPRC gain over endpoint; metric-dependent controls
SEA-RAFT/KITTI	simple convergence; richer-history null	tests added signal beyond native uncertainty
SPair-71k	timestep-aligned residual correspondence	alignment, not isolated chronology; wedge null
WikiText-103	distributed residual order	same executed head via endpoint-repeat; scale boundary

A.2 Proof of Proposition 1 and Local Nonlinear Chirality

Proof of Proposition 1. In superscripts such as $\big(R_\phi\big)^K$, parentheses denote repeated matrix multiplication, not transpose. Write $w_k(y) = z_k(y) - c$. Then $w_0 = v$ and $w_{k+1} = aR_{y\phi}w_k$, so

$$w_k(y) = a^k\big(R_{y\phi}\big)^k v, \tag{9}$$

which gives $z_k(y) = c + a^k\big(R_{y\phi}\big)^k v$. For an integer horizon K, the endpoint bound follows from

$$z_K(+1) - z_K(-1) = a^K\big((R_\phi)^K - (R_{-\phi})^K\big)v \tag{10}$$

and the triangle inequality.

Now define $u_y = (aR_{y\phi} - I)v$. Since $a < 1$, the matrix $aR_{y\phi} - I$ is nonsingular, so $u_y \neq 0$ whenever $v \neq 0$. We then have

$$z_{k+1} - z_k = w_{k+1} - w_k = a^k\big(R_{y\phi}\big)^k u_y, \tag{11}$$

so each summand in χ equals

$$\det(z_{k+1} - z_k,\ z_{k+2} - z_{k+1}) = a^{2k+1}\det(u_y,\ R_{y\phi}u_y). \tag{12}$$

For any nonzero $u \in \mathbb{R}^2$, planar rotation satisfies

$$\det(u,\ R_{y\phi}u) = y\sin(\phi)\|u\|^2. \tag{13}$$

Therefore

$$\chi(z_{0:K}) = y\sin(\phi)\|u_y\|^2 \sum_{k=0}^{K-2} a^{2k+1}, \tag{14}$$

and every factor except y is strictly positive. Hence $\operatorname{sign}(\chi(z_{0:K})) = y$ for every integer horizon $K \geq 2$. □

Proposition 2 (Local nonlinear chirality persistence)**.** *Let $F_y : \mathbb{R}^d \to \mathbb{R}^d$, $y \in \{-1, +1\}$, be C^2 maps with a common fixed point c. Let $E \subset \mathbb{R}^d$ be an oriented two-dimensional subspace with orthogonal projection P_E. Assume that, in a neighborhood $B_\rho(c)$, each F_y is a λ-contraction with $0 < \lambda < 1$, and that its Jacobian at c leaves E invariant with*

$$DF_y(c)|_E = aR_{y\phi}, \qquad 0 < a < 1, \quad 0 < \phi < \pi. \tag{15}$$

Fix a unit vector $v \in E$ and a finite integer horizon $K \geq 2$. For sufficiently small $\epsilon > 0$, initialize both classes at $z_0^y = c + \epsilon v$ and define $z_{k+1}^y = F_y(z_k^y)$. Then the endpoints collapse as

$$\|z_K^+ - z_K^-\| \leq 2\lambda^K \epsilon, \tag{16}$$

while the projected chirality functional

$$\chi_E(z_{0:K}^y) = \sum_{k=0}^{K-2} \det_E\big(P_E(z_{k+1}^y - z_k^y),\ P_E(z_{k+2}^y - z_{k+1}^y)\big) \tag{17}$$

satisfies $\operatorname{sign}(\chi_E(z_{0:K}^y)) = y$, where $\det_E$ denotes the oriented area form on E.

This proposition follows the same fixed-input convention as the main text: after the input/class is fixed, the same map F_y is reused across inner-time steps. The maps may differ across classes through the fixed input; the "shared-map" aspect is reuse over k for a given input, not class-agnostic dynamics with all label information only in the initial condition.

Proof. Write $w_k^y = z_k^y - c$. Since $F_y(c) = c$ and F_y is C^2, Taylor's theorem gives

$$w_{k+1}^y = A_y w_k^y + N_y(w_k^y), \qquad A_y = DF_y(c), \tag{18}$$

where $\|N_y(w)\| \leq C\|w\|^2$ in a small enough neighborhood of c. Contraction gives $\|w_k^y\| \leq \lambda^k \epsilon$, so trajectories remain in that neighborhood after reducing ϵ if necessary. Let $\bar{w}_k^y = A_y^k \epsilon v$. For every fixed K, induction on k gives

$$\max_{0 \leq k \leq K} \|w_k^y - \bar{w}_k^y\| \leq C_K \epsilon^2, \tag{19}$$

where C_K depends on the local C^2 bound and on K, but not on ϵ. Therefore the displacement errors obey

$$(w_{k+1}^y - w_k^y) - (\bar{w}_{k+1}^y - \bar{w}_k^y) = O_K(\epsilon^2) \tag{20}$$

uniformly over $0 \leq k < K$.

On E, the linearized dynamics reduce to the planar system:

$$P_E(\bar{w}_{k+1}^y - \bar{w}_k^y) = \epsilon a^k \big(R_{y\phi}\big)^k u_y, \qquad u_y = (aR_{y\phi} - I)v. \tag{21}$$

The exact linear contribution to χ_E is

$$y\, \epsilon^2 \sin(\phi) \|u_y\|^2 \sum_{k=0}^{K-2} a^{2k+1}. \tag{22}$$

Because determinant is bilinear, replacing the exact linear displacements by the nonlinear displacements changes the finite sum by at most $C'_K \epsilon^3$ for another finite constant C'_K. The coefficient of ϵ^2 above is strictly positive apart from the class sign y. Hence, for all sufficiently small ϵ, the $O_K(\epsilon^3)$ nonlinear remainder cannot flip the sign, and $\operatorname{sign}(\chi_E(z_{0:K}^y)) = y$.

Finally, since c is a common fixed point and each F_y is a λ-contraction,

$$\|z_K^+ - z_K^-\| \leq \|z_K^+ - c\| + \|z_K^- - c\| \leq 2\lambda^K \epsilon. \tag{23}$$

□

This proposition is a local nonlinear sufficient condition for the same phenomenon as Proposition 1. It is not a characterization of arbitrary learned update maps: the assumptions require a common attractor, a contracting neighborhood, and an oriented two-dimensional mode in the Jacobian. Within that setting, however, the result is genuinely nonlinear and non-telescoping: the signed-area functional depends on consecutive triples of states, and the C^2 nonlinear remainder is controlled rather than ignored.

When the assumptions weaken. If the two contractive maps have distinct attractors c_+ and c_-, the triangle inequality gives, while each trajectory stays in its contraction neighborhood,

$$\|z_K^+ - z_K^-\| \le \|c_+ - c_-\| + \lambda^K \big(\|z_0^+ - c_+\| + \|z_0^- - c_-\|\big).$$

If both initial distances are at most ϵ, the second term is at most $2\lambda^K\epsilon$. Attractor mismatch therefore leaves a non-decaying endpoint-separation term; it need not destroy chirality, but removes the guarantee that endpoints collapse together. The two-dimensional mode is a subspace condition, not a restriction to two-dimensional networks. For a fixed horizon, write the linear signed-area margin as $m_y\epsilon^2$ with $m_y > 0$ from the proof. Its sign is protected only while the total signed-area perturbation is smaller than that margin; the local nonlinear bound suffices when $C'_K\epsilon^3 < m_y\epsilon^2$. A missing oriented mode, coupling outside the controlled neighborhood, or a larger remainder removes that guarantee. This is a sufficient-condition witness, not a prediction of chirality in arbitrary learned systems.

A.3 Reproducibility: Readout, Resource, and Asset Details

Trajectory readouts. For reproducibility, the controlled endpoint-collision construction uses $z_{i,0} = c_i + 1$ and $z_{i,k+1} = a_i z_{i,k} + (1 - a_i)c_i$ with $c_i \sim \mathrm{Uniform}[-3, 3]$, $a_0 = 0.2$, $a_1 = 0.9$, and $K = 40$. Its latent state is one-dimensional, but its settling-history readout is the four-feature aggregate vector $[z_T, \mu(\Delta z), \mathrm{var}(\Delta z), \sum_k \|\Delta z_k\|_2]$, where $\Delta z_k = z_{k+1} - z_k$. The learned frozen-probe collision study and the shared-map ablation use the same form, with the shuffled control keeping z_T fixed and reassigning only the refinement-history-summary part from another example in the same split. The binned learned-collision run uses a slightly richer compact summary,

$$[z_T,\ z_0,\ \mu_k\|\Delta z_k\|_2,\ \mathrm{var}_k\|\Delta z_k\|_2,\ \sum_k \|\Delta z_k\|_2,\ \mu_k z_k,\ \mathrm{var}_k z_k].$$

Endpoint-strength controls always receive endpoint-derived features only. They use a fixed map $\psi(z_T)$ followed by the same linear/logistic classifier used for the history features. In the controlled and learned collision runs, ψ is a label-blind endpoint replication or random projection chosen so that $\dim \psi(z_T)$ matches the trajectory feature dimension; this makes the fitted linear classifier's input size match the corresponding trajectory probe without introducing any learned hidden layer or non-endpoint input. In the frozen-backbone and endpoint-capacity controls, the stronger endpoint control uses a fixed nonlinear expansion such as $[z_T, z_T^2, \sin(z_T), \|z_T\|_2]$ before the same linear/logistic classifier. We therefore use these controls to test feature-count and endpoint-readout-strength loopholes, not to claim that a single linear layer on raw low-dimensional z_T can match the parameter count of a larger history-feature probe. For the path-functional probe in Section 5.2, the shared-map trajectory probe uses the frozen feature vector

$$[z_T,\ \mathrm{vec}(\mu_k(\Delta z_k)),\ \{\sum_k \|\Delta z_k\|_2\}_{\mathrm{outer}},\ \{\|\Delta z_{T-1}\|_2\}_{\mathrm{outer}}],$$

where $\Delta z_k = z_{k+1} - z_k$ denotes inner-refinement steps and "outer" indexes the synthetic path positions. We also include an endpoint-sequence baseline that flattens the final inner endpoint at every outer position, $(z_{1,T}, \ldots, z_{64,T})$, but excludes inner-step displacements, path lengths, and residual features. This baseline isolates whether the trajectory probe is winning only because it sees endpoint values across external-time sequence positions. The raw-path leak controls are separate: endpoint, final delta, last-five window, simple raw-path statistics $[u_T, u_0, \mu(u), \sigma(u), u_T - u_{T-1}, u_T - u_{T-5}, \min(u), \max(u)]$, and a path-geometry oracle using sign-change rate and mean absolute second difference. For the topology-conflict visual stress test, the compact reconstruction-history summary readout is built from per-step pooled latent statistics of the reconstruction state: the spatial mean and standard deviation of each latent channel at every refinement step, concatenated with their first differences across steps. The endpoint leak controls are kept separate: final reconstruction error, final delta norm, and simple endpoint-state summary statistics.

Compute and storage. The controlled probe-suite experiments use small synthetic datasets, fixed declared feature maps with linear/logistic probes, and compact recurrent backbones; they run on a single workstation without distributed training. Some fixed summaries, including running path length and update moments, can be accumulated with $O(d)$ extra memory. This does not extend

automatically to concatenated states, endpoint-relative histories, or learned full-history heads. Full-path caching costs $O(Td)$ and supports frozen audits, shuffling, residualization, and reproducibility checks. The heaviest reported applied stress test is the topology-conflict visual probe: we generate a 3200-example tiled visual dataset, train small iterative reconstructors on CPU, and evaluate frozen probes on full and endpoint-boundary subsets. The standard checks require external assets: SPair-71k diffusion features and RAFT/KITTI trajectory caching are GPU-assisted, while the reported probe summaries are small result-summary files. The reported results are compact and do not require storing large auxiliary sweeps in the supplementary material.

Broader impacts. Endpoint-sufficiency audits can expose failure information hidden by an apparently stable prediction. Their potential benefit is more informative reliability evaluation; the corresponding risk is over-trusting post-hoc scores in safety-critical systems. These experiments do not validate such deployment, and domain-specific testing remains necessary.

Existing assets. The controlled probe suite and topology-conflict visual stress test use locally generated data. The standard-system checks use public third-party assets: KITTI-2015 and pretrained RAFT/SEA-RAFT for optical flow, SPair-71k for semantic correspondence, WikiText-103 for language modeling, and Diffusion Hyperfeatures/Stable Diffusion feature extraction for the diffusion audit. These assets are cited in the main text and are not redistributed in the supplementary material. WikiText is distributed under CC BY-SA 3.0 and GFDL terms; Diffusion Hyperfeatures code and released aggregation weights are Apache-2.0; Stable Diffusion model use follows the CreativeML Open RAIL-M terms; KITTI and RAFT are used under their public research benchmark/checkpoint terms; and SPair-71k is used under its public research benchmark terms, including the underlying PASCAL/VOC and Flickr source-image terms noted by the dataset. Reproducing the standard-system checks therefore requires obtaining those assets from their original providers and respecting their licenses or terms of use.

A.4 Shared-Map Ablation Details

Figure 2 summarizes the shared-map mechanism diagnostics; the paragraphs below give the update equations, hard-slice comparisons, and sweep controls.

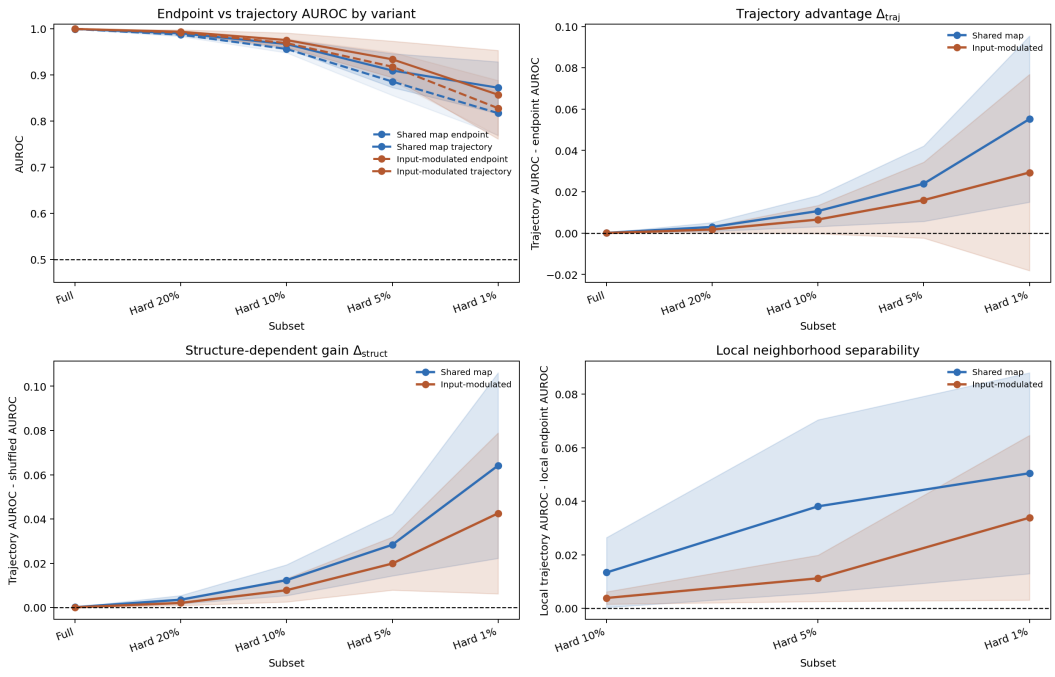


Figure 2: **Input-modulation diagnostics.** Both variants retain time-tied updates. Modulation reduces observed trajectory gains on variant-specific slices; this does not isolate shared-weight causality.

Shared-map ablation. This ablation tests sensitivity to input-dependent modulation, not removal of weight tying. Both variants instantiate the fixed-input, time-tied update $z_{k+1} = F_\theta(z_k, x)$ from the main text. The baseline shared-map update is

$$F_\theta^{\text{shared}}(z_k, x) = \tanh(W_z z_k + W_x x),$$

while the input-modulated variant is

$$F_{\theta,\alpha}^{\text{mod}}(z_k, x) = \tanh(W_z z_k + m_\alpha(x) \odot z_k + W_x x),$$

where x is the fixed input and $m_\alpha(x) = \alpha \tanh(Mx + b_m)$ is an input-conditioned diagonal perturbation that makes the update more input-specific. The scalar α is the modulation-strength sweep parameter in Figure 3; the main input-modulated comparison uses $\alpha = 0.06$, and the sweep varies $\alpha \in \{0, 0.02, 0.05, 0.10\}$. The two backbones remain matched on the original endpoint task (endpoint-head AUROC 0.9995 ± 0.0001 for the shared model versus 0.9997 ± 0.0001 for the input-modulated model), so the intervention does not simply weaken the classifier. For compactness in the diagnostics below, let E denote the endpoint-only probe, T the refinement-history summary probe, and S the shuffled-history control.

The primary quantity is the trajectory advantage $\Delta_{\text{traj}} = \text{AUROC}(T) - \text{AUROC}(E)$. On nominal top 5% and 1% collision-ranked subsets, Δ_{traj} is 0.0239 and 0.0552 for the unmodulated map, versus 0.0160 and 0.0293 for the modulated variant. The corresponding normalized reductions are 33.2% (95% CI $[16.5\%, 93.9\%]$) and 46.9% (95% CI $[13.0\%, 132.5\%]$). These are 600-draw paired-seed bootstrap intervals over five seeds; ratio intervals are unstable when the denominator is small. Each model constructs its own label-informed collision ranking, so slices need not contain the same examples. The nominal Hard 1% slice is floored at 50 of 2,500 cases (2%). This is a retrospective diagnostic, not deployable label-free selection.

The same pattern appears in two corroborating diagnostics. First, the task-aligned structure control $\Delta_{\text{struct}} = \text{AUROC}(T) - \text{AUROC}(S)$ is larger for the unmodulated map on Hard 5% (0.0284 versus 0.0199; difference 0.0085, 95% CI $[0.0015, 0.0132]$) and nominal Hard 1% (0.0642 versus 0.0426), though the latter is noisier. Second, within collided z_T neighborhoods, the local AUROC gain is 0.0504 versus 0.0338 on nominal Hard 1%. These outcomes support sensitivity to input modulation. Because modulation adds parameters, changes learned dynamics, and changes the evaluated collision neighborhoods, they do not isolate shared weights as the causal explanation.

Table 5: **Input-modulation diagnostics.** Both maps are time-tied. Nominal Hard 1% contains 50 cases; rankings are variant-specific. These results are supportive, not an isolated causal test of weight sharing.

Variant	Hard 5% Δ_{traj}	Hard 1% Δ_{traj}	Hard 5% Δ_{struct}	Hard 1% local gap
Shared map	0.0239	0.0552	0.0284	0.0504
Input-modulated	0.0160	0.0293	0.0199	0.0338

The main text already shows the core task-aligned mechanism curves. The appendix keeps only the sweep figure below, which extends the ablation beyond the single main-text comparison.

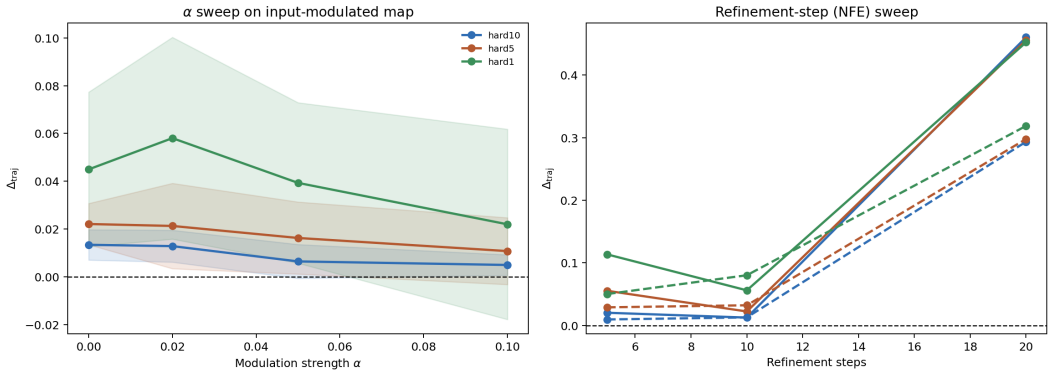


Figure 3: **Sweep diagnostics.** Left: fixed-seed α sweep on the input-modulated variant. Increasing modulation strength reduces Δ_{traj} monotonically on Hard 10% and Hard 5%; the Hard 1% curve is noisier, though its lowest observed value is still at $\alpha = 0.10$. Right: compute-matched refinement-step (NFE) sweep, where NFE denotes the number of function evaluations. In the right panel, color encodes the subset and line style encodes the variant (solid = shared map, dashed = input-modulated). The shared-map model remains above the input-modulated variant at 5 and 20 inner steps on the hardest subsets, while the intermediate 10-step point is mixed, so we treat the refinement-step sweep as supporting but secondary evidence.

A.5 Path-Functional Collapse Details

Path-functional classification under endpoint collapse. Each sample is a 64-position synthetic path with independent binary stable and path labels. Backbones train only on the stable nuisance label, which all solve perfectly; frozen probes then predict the path label. This deliberately separates the training task from the audit target. It does not rule out a recurrent model trained to store the path label as a persistent memory bit. The stronger endpoint-sequence probe receives $(z_{1,T}, \ldots, z_{64,T})$, but no inner-step displacement, path-length, or residual summaries.

Generation and input access. The first 54 positions integrate 53 noisy signed increments. Class zero uses alternating sign runs of uniformly sampled length 5–10; class one alternates every increment, with random initial phase and independent sign slips of probability 0.08. Increment magnitudes are log-normal with log mean -0.2 and standard deviation 0.25, plus Gaussian noise of standard deviation 0.10. A noisy ten-position landing segment ends at zero. The normalization first subtracts the endpoint-interpolating line, removes the sample mean using a zero-endpoint quadratic bump, divides by standard deviation, and re-subtracts the endpoint line. Gaussian observation noise of standard deviation 0.04 is added and the same normalization applied again. The stable channel is its signed label plus Gaussian noise of standard deviation 0.03. There are 5,000/1,500/2,500 training/validation/test cases per seed, three seeds, 50 backbone epochs, and two backbone restarts. The released generator specifies the exact landing construction and numerical operations.

The shared-map backbone receives the path input only at the first of eight inner updates at each external position; subsequent updates use the stable channel alone. GRU receives both channels at each external position. DEQ receives only the stable channel throughout, so its chance path-label result is a no-access control, not evidence of information loss during equilibrium solving. These are not equal-access architectural comparisons. The transient-injection schedule also differs from a stationary fixed-input map over the entire inner episode.

Subset and uncertainty qualifications. Normalization makes every raw path endpoint exactly zero. The recorded endpoint bands sort these tied values; they are reproducible tie-selected subsets, not graded endpoint-collapse difficulty. The nominal top-1% band contains 80 cases (the minimum), not 25. Full-set results are the primary comparison. The saved aggregate gap ranges are the 2.5th and 97.5th percentiles of three observed seed gaps, not pooled bootstrap draws and not 95% confidence intervals for the mean. Per-seed example bootstraps exist separately but do not justify that label for the aggregate ranges.

Stable-task AUROC is 1.0000 ± 0.0000 for all three backbones. The final shared-map endpoint reaches 0.4932 ± 0.0065, DEQ endpoint 0.5013 ± 0.0060, and the raw-path shortcut probes give 0.5000 for endpoint, 0.4833 for final delta, 0.4890 for the last five steps, and 0.5367 for simple statistics. Endpoint-sequence reaches 0.6682 ± 0.1386, versus shared-map trajectory 0.9180 ± 0.0337 and shuffled 0.4981 ± 0.0049. The trajectory-minus-sequence mean gap is 0.2498, with descriptive seed-quantile range $[0.1099, 0.3455]$. GRU endpoint reaches 0.5553 ± 0.0634; the cross-backbone gap is 0.3626 with descriptive range $[0.2727, 0.4253]$. Gaps against shared-map endpoint, shuffled control, and DEQ are 0.4248, 0.4198, and 0.4167. The trajectory also scores 0.9275 and 0.9247 on the nominal top-10% and top-1% subsets, respectively, subject to the tie qualification above.

The strongest evidence is the same-backbone comparison: history exposes the audit target better than the tested endpoint and shortcut decoders. This is neither an impossibility result for stronger endpoint decoders nor a claim that trajectory readouts dominate recurrent memory trained on the path label. No conclusion about horizon-independent performance follows from the single 64-position setting.

The main text already shows the core frozen-readout and leak-control curves. The table below records the exact endpoint-band summary values used for the appendix discussion.

Table 6: **AUROC summary.** Frozen probes on the full set and tied-endpoint subsets. Nominal band labels are retained for artifact correspondence, not hardness claims; the 1% band is floored at 80 cases. DEQ never receives the path channel.

Readout	Full	Top 10% endpoint band	Top 1% endpoint band
Shared-map endpoint	0.4932	0.4811	0.5312
Shared-map endpoint sequence	0.6682	0.6977	0.6779
Shared-map trajectory	0.9180	0.9275	0.9247
Shuffled control	0.4981	0.5098	0.5047
DEQ endpoint	0.5013	0.5072	0.5108
GRU endpoint	0.5553	0.5570	0.5382

A.6 Endpoint-Residualized Synthetic Audits

Setup recap. Section 6 defines train-only ridge residualization, followed by ordered and permutation-augmented probes. For the path-functional probe, the residualized object is external-position endpoint history $(z_{1,T}, \ldots, z_{64,T})$ conditioned on $z_{64,T}$, not the inner-step summary of the main trajectory probe. Collision and topology-conflict audits use inner-refinement states. A high-capacity nonlinear endpoint MLP is also trained directly on $z_{64,T}$ for the path-functional probe. Paired gaps in Tables 8–10 are thresholded classification-accuracy differences, not AUROC differences. Their brackets average the lower and upper endpoints of within-seed example-bootstrap intervals. Such averages are descriptive mean within-seed bounds, not 95% confidence intervals for the across-seed mean; they do not include seed-level uncertainty.

Per-step endpoint predictivity. The residualized audit is sharpest when the reference endpoint already explains most of the raw ordered history. Table 7 shows that all three settings satisfy this condition under the train-only ridge fits. Thus any persistent ordered-versus-permuted gap on the residual reflects structure that survives this linear endpoint-conditioning control.

Table 7: Endpoint predictivity before residualization. The endpoint linearly explains most raw ordered history in all three residual-audit settings.

Setting	Ridge predictivity target	Mean R^2 range
Path-functional probe	$R^2(z_{j,T} \mid z_{64,T})$ across 64 external positions	$[0.9947, 0.9985]$
Collision-severity ladder	$R^2(z_t \mid z_T)$ across 12 inner-time steps	$[0.911, 0.994]$
Topology-conflict visual probe	$R^2(z_t \mid z_T)$ across five inner-refinement steps	$[0.986, 0.997]$

Path-functional probe. Table 8 reports the full set and tie-selected endpoint bands of **Appendix A.5**. All raw endpoints are zero, so the bands do not establish a severity trend. The nominal top-1% slice is floored at 80 cases. Ordered residual accuracy exceeds both controls, with substantial seed variability in AUROC.

Table 8: Path-functional residual audit. Accuracy triples are thresholded accuracies for endpoint MLP, ordered residual, and permutation-augmented residual readouts.

Slice	n	Endpoint MLP AUROC	Ordered residual AUROC	Permuted residual AUROC	Accuracy endpoint/ordered/permuted	Ordered − endpoint accuracy gap	Ordered − permuted accuracy gap
Full	2500	0.499 ± 0.012	0.863 ± 0.097	0.509 ± 0.005	$0.502/0.797/0.495$	$0.295\ [0.271, 0.320]$	$0.302\ [0.278, 0.326]$
Top 20%	500	0.503 ± 0.027	0.861 ± 0.100	0.510 ± 0.013	$0.490/0.801/0.507$	$0.311\ [0.256, 0.366]$	$0.294\ [0.241, 0.347]$
Top 10%	250	0.479 ± 0.021	0.868 ± 0.094	0.480 ± 0.024	$0.488/0.815/0.493$	$0.327\ [0.249, 0.404]$	$0.321\ [0.245, 0.395]$
Top 5%	125	0.456 ± 0.080	0.861 ± 0.099	0.446 ± 0.037	$0.501/0.808/0.485$	$0.307\ [0.197, 0.411]$	$0.323\ [0.213, 0.429]$
Top 1% floor	80	0.473 ± 0.068	0.887 ± 0.082	0.467 ± 0.056	$0.525/0.842/0.500$	$0.317\ [0.188, 0.442]$	$0.342\ [0.208, 0.471]$

Collision-severity ladder. Table 9 reports the collision ladder of Section 5. Mean ordered-residual performance exceeds the permutation control. The nominal Hard 1% slice is floored at 50 examples and has wide mean within-seed bounds; these are not across-seed significance tests.

Table 9: Collision-severity residual ladder. Accuracy pairs are thresholded accuracies for ordered residual and permutation-augmented residual readouts.

Slice	n	Ordered residual AUROC	Permuted residual AUROC	Accuracy ordered/permuted	Ordered − permuted accuracy gap
Full	2500	0.998 ± 0.001	0.841 ± 0.051	$0.985/0.778$	$0.207\ [0.191, 0.223]$
Hardest 20%	500	0.990 ± 0.008	0.789 ± 0.056	$0.952/0.737$	$0.215\ [0.175, 0.255]$
Hardest 10%	250	0.970 ± 0.021	0.719 ± 0.029	$0.909/0.700$	$0.209\ [0.143, 0.272]$
Hardest 5%	125	0.923 ± 0.059	0.664 ± 0.057	$0.845/0.653$	$0.192\ [0.096, 0.291]$
Hardest 1% floor	50	0.876 ± 0.084	0.610 ± 0.088	$0.773/0.613$	$0.160\ [-0.007, 0.327]$

Topology-conflict visual probe. Table 10 is a residual-null result for the topology-conflict probe (**Appendix A.11**). Both tested readouts are near chance and all descriptive mean within-seed bounds include zero. High endpoint predictivity and the residual null are consistent with endpoint-recoverability of the raw gain. They do not establish absence of label information in residuals or sufficiency of the endpoint: a limited probe may fail to recover information that remains.

Table 10: Topology-conflict residual audit. Accuracy pairs are thresholded accuracies for ordered residual and permutation-augmented residual readouts.

Slice	n	Ordered residual AUROC	Permuted residual AUROC	Accuracy ordered/permuted	Ordered − permuted accuracy gap
Full	1000	0.484 ± 0.008	0.485 ± 0.019	$0.489/0.479$	$0.011\ [-0.027, 0.048]$
Hardest 50%	500	0.485 ± 0.010	0.482 ± 0.004	$0.493/0.473$	$0.019\ [-0.035, 0.072]$
Hardest 25%	250	0.480 ± 0.010	0.469 ± 0.021	$0.491/0.463$	$0.028\ [-0.049, 0.105]$
Hardest 10%	100	0.447 ± 0.040	0.458 ± 0.024	$0.457/0.437$	$0.020\ [-0.110, 0.150]$

Reading the audit. A positive residual comparison identifies signal surviving the specified affine conditioning and probe controls. A null fails to demonstrate such signal; it does not prove its absence. Neither result settles all nonlinear endpoint decoding possibilities. We report both outcomes and the tested probe family.

A.7 Orientation-Aware Trajectory Audit

Orientation-aware trajectory audit. To connect the signed-area witness in Proposition 1 to the empirical readouts, we compute two high-dimensional orientation-aware path features: a local wedge $\sum_k \Delta z_k \wedge \Delta z_{k+1}$ and the antisymmetric level-2 signature / Lévy-area feature $\sum_{\ell<k} \Delta z_\ell \wedge \Delta z_k$. Table 11 reports the full-split audit. Local-wedge and Lévy-area features are far above endpoint and step-permuted controls; the same pattern survives endpoint residualization. Testing a forward-trained local-wedge probe on time-reversed trajectories drops below chance, consistent with the antisymmetric sign flip. These numbers test whether the theorem's orientation-aware witness is also an empirical mechanism, rather than only an existence proof.

Table 11: Orientation-aware trajectory audit. Step permutation destroys orientation-aware signal, and time reversal flips the sign-sensitive readouts.

Readout / control	AUROC	Role
Endpoint	0.4932 ± 0.0065	endpoint-only
Endpoint sequence	0.6685 ± 0.1378	endpoint-history control
Magnitude summary	0.9179 ± 0.0337	non-oriented path magnitude
Local wedge	0.9733 ± 0.0018	orientation-aware path feature
Local wedge, step-permuted	0.5113 ± 0.0073	order-destroying control
Local wedge, time-reversed	0.1426 ± 0.1652	reversal-sign check
Lévy area	0.9729 ± 0.0015	level-2 antisymmetric feature
Lévy area, step-permuted	0.5077 ± 0.0026	order-destroying control
Lévy area, time-reversed	0.1801 ± 0.2177	reversal-sign check
Residual local wedge	0.9214 ± 0.0110	after endpoint conditioning
Residual local wedge, step-permuted	0.5040 ± 0.0074	residual order control
Residual Lévy area	0.9297 ± 0.0109	after endpoint conditioning
Train-selected scalar wedge coordinate	0.7262 ± 0.0246	single-coordinate sign check

The local-wedge minus step-permuted gap is $+0.4620$ with descriptive mean within-seed bounds $[+0.4377, +0.4855]$; the residualized comparison is $+0.4174$ with bounds $[+0.3939, +0.4418]$. These average per-seed bootstrap interval endpoints, not bootstrap draws across seeds, and are not confidence intervals for the across-seed mean. The AUROC standard deviations above separately describe seed variability.

A.8 Compute-Matched Path-Functional Control Details

Compute-matched path-functional control. This control compares endpoint and history access on the same trained backbone at update budgets $K \in \{2, 4, 8, 16\}$. The shared-map model performs K refinement steps, DEQ is capped at K solver evaluations, and GRU receives K recurrent updates per position. Equal update counts do not equate cell FLOPs, parameters, or probe overhead. The same-backbone comparison needs no extra upstream updates for history. As in **Appendix A.5**, DEQ receives no path input, and raw endpoint bands are tie-selected subsets rather than a difficulty ranking.

The compute-matched result preserves the main separation (Table 12). On the full test set, the shared-map trajectory remains the strongest readout at every budget, while endpoint-only and baseline readouts remain near chance. On the top-10% endpoint band, the shared-map trajectory shows the same pattern. Shortcut probes remain controlled across budgets: endpoint 0.5000, final delta 0.4930, last-five-step 0.4938, and simple statistics 0.5231, whereas the path-geometry oracle (a probe given geometry-aware features computed from the whole path) is 1.0000. Thus the path-functional gain is not explained by extra upstream compute; it comes from representational access to the already-computed refinement path.

Table 12: Compute-matched path-functional control. At each budget, the trajectory readout only accesses states already computed by that fixed-budget rollout.

K	Shared endpoint full AUROC	Shared trajectory full AUROC	Shared shuffled full AUROC	DEQ endpoint full AUROC	GRU endpoint full AUROC	Shared trajectory top-10% AUROC
2	0.5347 ± 0.0128	0.9696 ± 0.0049	0.5165 ± 0.0058	0.5093 ± 0.0068	0.5057 ± 0.0111	0.9695 ± 0.0192
4	0.5056 ± 0.0020	0.9550 ± 0.0019	0.5106 ± 0.0031	0.5068 ± 0.0100	0.5051 ± 0.0029	0.9583 ± 0.0174
8	0.5067 ± 0.0043	0.9287 ± 0.0212	0.5178 ± 0.0040	0.5006 ± 0.0113	0.5043 ± 0.0013	0.9272 ± 0.0476
16	0.5067 ± 0.0043	0.7881 ± 0.0729	0.5083 ± 0.0125	0.5003 ± 0.0111	0.5024 ± 0.0031	0.7450 ± 0.0902

Performance decreases at $K = 16$, but this does not identify erasure from the full history. The trained backbone and finite summary change with the budget; a longer raw history on a fixed rollout still contains its earlier prefix. The supported statement is that this trained summary/probe pipeline is not monotonically better with more updates.

The main text already shows the full-set matched-budget curve. The appendix keeps only the endpoint-band and shortcut-probe details that extend that result.

The shortcut panel tests final value, last-step change, a short endpoint-adjacent window, and simple global statistics. These readouts do not solve the task, whereas a whole-path geometry probe does. This excludes the tested shortcuts, not every possible compact statistic or learned persistent memory.

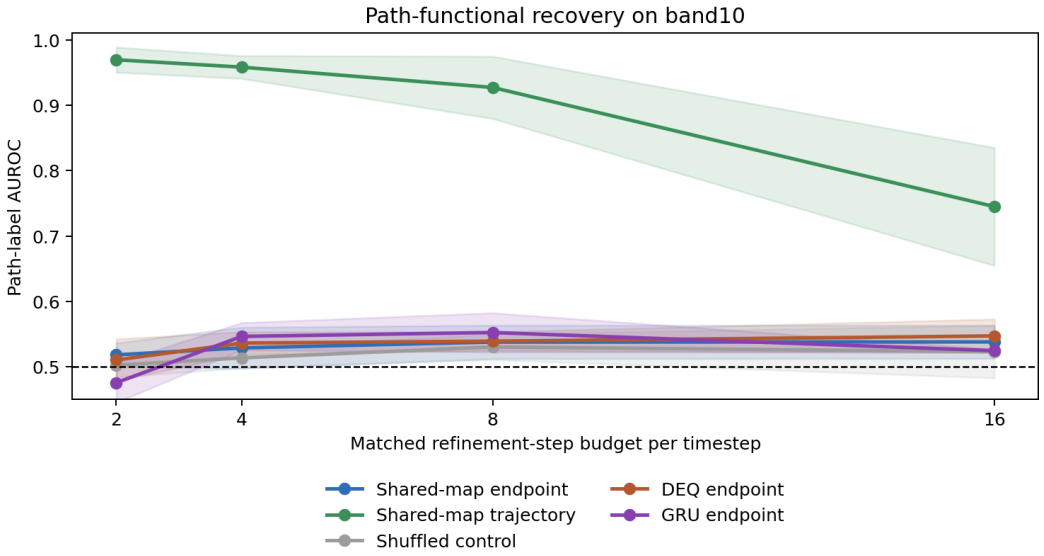


Figure 4: **Tie-selected endpoint subset.** The nominal top-10% band is not a hardness ranking because raw endpoints tie at zero. Performance of this summary/readout declines at $K = 16$; this does not establish erasure from full history.

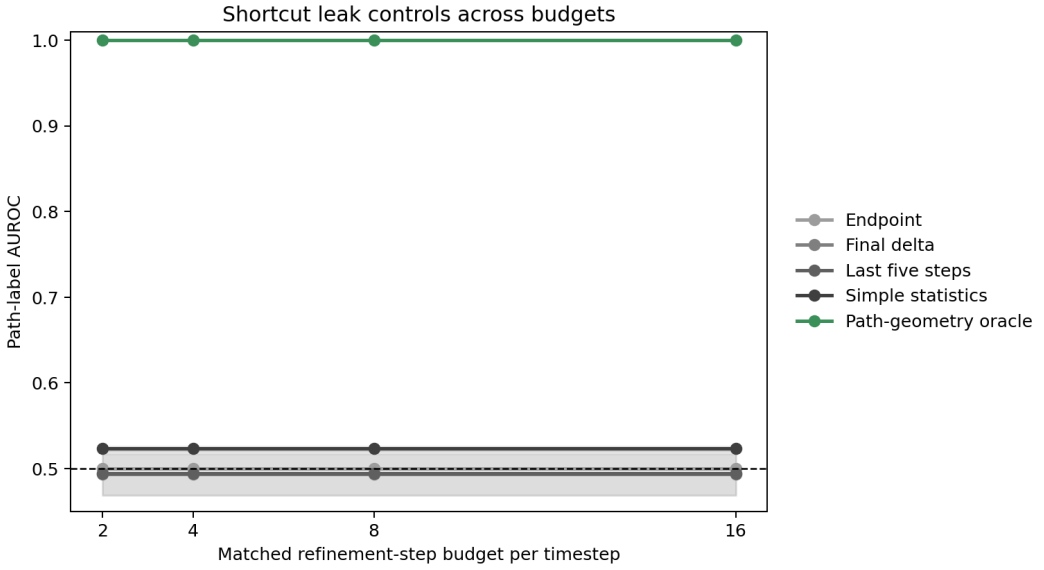


Figure 5: **Shortcut probes.** The tested endpoint, final-delta, short-window, and simple-statistic probes are weak, whereas the whole-path geometry probe is perfect. This is an update-budget control, not FLOP- or latency-matched architectural superiority.

A.9 Cross-Model Transient-Washout Details

This subsection is the appendix target for the cross-model washout scope check in Section 5.3. Each example contains $x = (s, t, n_1, n_2)$ with $s, t, n_1, n_2 \sim \mathcal{N}(0, 1)$. Backbones train only on the stable

label $y_{\text{stable}} = \mathbf{1}[s > 0]$. After training, the backbone is frozen and linear probes test the transient label $y_{\text{trans}} = \mathbf{1}[t > 0]$. Hard washout subsets are ranked by $|s|$ on the test split: larger $|s|$ means a stronger stable drive and a harder endpoint-only transient probe.

The shared-map and GRU backbones receive the same stable/transient separation. Let $u_s = E_s(s, n_1, n_2)$ and $u_t = E_t(t, n_1, n_2)$. The shared-map update is

$$z_0 = 0, \qquad z_{k+1} = \tanh\big(W_z z_k + W_x\big(u_s + \mathbf{1}[k = 0]u_t\big)\big), \qquad k = 0, \ldots, K-1.$$

The GRU baseline uses the same input schedule with a GRUCell. The DEQ endpoint baseline solves the stable-drive fixed point

$$z^\star = \tanh(W_z z^\star + W_x u_s),$$

so it never observes the transient cue and is a no-access control, not evidence that the equilibrium solver erased information it received. All three backbones reach stable-task AUROC 1.0000 ± 0.0000 before the frozen transient probes are fit. The main-text numbers report endpoint-only probes for shared-map, DEQ, and GRU, plus the shared-map refinement-history summary probe; the GRU refinement-history summary reaches 0.6817 ± 0.0981 AUROC on Hard-1%, versus its endpoint at 0.5317 ± 0.0692. This is a scope check, not a general architectural superiority claim.

A.10 Linear SSM Backbone Control Details

Additional backbone control: linear SSM on the path-functional task. To test whether the endpoint–trajectory gap is specific to the shared-map backbone, we rerun the path-functional protocol with a minimal contractive linear state-space model (SSM). The backbone is trained only on the stable nuisance cue and then frozen. The linear SSM endpoint probe uses the same frozen endpoint-only readout protocol as before, trained on the final state alone. We fit the same endpoint-only, state-trajectory summary, and shuffled-history probes to predict the path-functional label. On the full test set, the endpoint probe reaches 0.5873 AUROC, while the trajectory probe reaches 0.9995; the shuffled control drops to 0.5750. Thus the endpoint-collapse phenomenon is not unique to one model family: even a simple linear state-space backbone can carry recoverable path signal in its state trajectory, while its final endpoint remains weak.

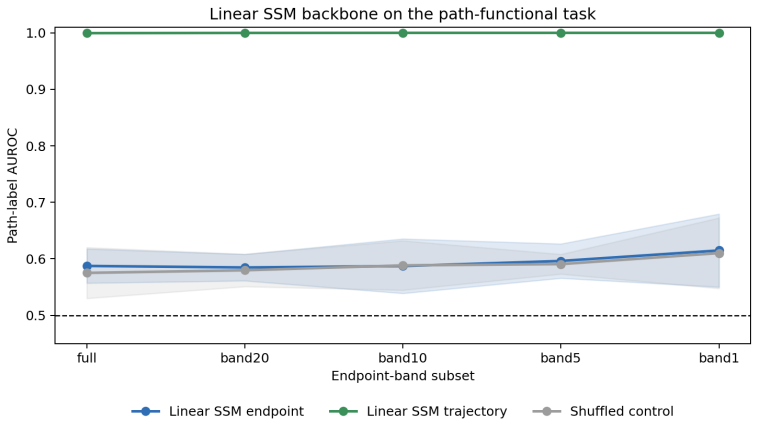


Figure 6: **Linear SSM backbone control.** Frozen endpoint-only, state-trajectory summary, and shuffled-history controls on the path-functional task.

A.11 Topology-Conflict Visual Stress Test

This appendix-only stress test is a residual-null boundary, not positive irreducibility evidence. The unconditioned trajectory readout beats endpoint probes, but tested residual probes do not recover the gain. This is consistent with endpoint-recoverability, not proof of it (Section 6).

Topology-conflict visual stress test. To test the endpoint-insufficiency claim on a fixed-input visual task without relying on coordinate-defined route labels, we introduce a topology-conflict

stress test in which both classes are corrupted away from the same underlying target image. Class 0 applies a coherent row/column cyclic shift, while class 1 applies a conflicting nonlocal three-tile cycle; both are then reconstructed back to the original target under a frozen-probe protocol. This keeps final reconstruction difficulty closely matched while varying the structural consistency of the refinement problem. On the full test distribution, the shared-map backbone reaches endpoint AUROC 0.649 ± 0.014 and trajectory AUROC 0.701 ± 0.028, while simple shortcut controls remain near chance (final reconstruction error 0.525 ± 0.021, final-delta norm 0.524 ± 0.031). The more informative view is the endpoint-boundary ladder: on the shared-map hard-50% slice, endpoint drops to 0.545 ± 0.023 while trajectory remains at 0.649 ± 0.021 and shuffled trajectory stays at 0.527 ± 0.028, increasing the ordered-trajectory gap from $+0.052$ on the full set to $+0.104$. DEQ and GRU reconstructors show the same endpoint-to-trajectory ordering, with gaps also widening on endpoint-boundary slices. Thus, under structural contradiction rather than external-time memory, trajectory readouts become more useful exactly where endpoint sufficiency weakens.

The endpoint-residualized audit changes the interpretation. Ordered residual AUROC is 0.484 ± 0.008, versus 0.485 ± 0.019 for permutation-augmented residuals. Both tested readouts are near chance. Endpoint-predictable visual modes could explain the raw gain by giving the probe easier intermediate views, but this is an interpretation, not an identified mechanism. A failed residual probe does not show that all useful history is linearly endpoint-recoverable. This is a residual-null scope check, not positive irreducibility evidence or proof of sufficiency.

Table 13: **Topology-conflict degeneracy scaling.** Mean AUROC gap between ordered trajectory and endpoint-only probes across slices where endpoints increasingly overlap. All three reconstructors show positive endpoint-to-trajectory gaps, and the gaps generally increase on endpoint-boundary slices. The hard-5% slice is the noisiest (smallest per-seed n), so the largest small-slice gaps should be read as directional stress-test evidence rather than a stable ranking of backbones.

Backbone	Full gap	Hard50	Hard25	Hard10	Hard5
Shared-map	+0.052	+0.104	+0.093	+0.089	+0.067
DEQ	+0.048	+0.101	+0.094	+0.127	+0.093
GRU	+0.053	+0.066	+0.101	+0.109	+0.148

A.12 RAFT/KITTI Convergence Reliability

Table 14: **Corrected RAFT/KITTI reliability.** Means of three seed-specific pooled metrics; AUPRC uses average precision and EPE is in pixels. Aggregate convergence is primary. Higher ranking scores and lower EPE are better. Paired physical-pair intervals are in Appendix A.12.

Readout	AUROC	AUPRC	EPE at 80%
Strict endpoint	0.8676	0.1399	0.4007
Endpoint + last update	0.8781	0.1881	0.3961
Variance + update energy	0.8911	0.1416	0.3072
Endpoint-overmatch MLP	0.9035	0.2128	0.3988
Aggregate convergence	0.9131	0.3143	0.3772
Ordered convergence	0.9073	0.3017	0.3788
Shuffled convergence	0.9009	0.2849	0.3706
Endpoint-displacement features	0.9111	0.2612	0.3022

Correction and evaluation population. An earlier adapter called RAFT's feature/context encoders directly and omitted the RGB transformation $2(I/255) - 1$ normally performed by `forward`. The corrected cache applies that transformation before either encoder. This changes the flow predictions, histories, and failure labels of all readouts; it is not a unit conversion. The old RAFT tables and old cache-rebuild/sanity-check numbers are superseded, not validation of this run.

We retain the public KITTI-finetuned checkpoint, KITTI-2015 training images, 24 iterations, and seeds 0, 1, and 2. Each reliability-probe split has 130 training pairs and 70 held-out pairs. Test pixels are sampled uniformly from valid ground-truth pixels, 4,096 per pair, not balanced by class.

These pairs are held out for the reliability probes, not the pretrained RAFT checkpoint. The outlier label is $\mathbf{1}[\mathrm{EPE} > 3\ \mathrm{px} \ \wedge\ \mathrm{EPE}/\max(\|u_{\mathrm{GT}}\|, 10^{-6}) > 0.05]$. Mean sampled held-out prevalence is 1.7980%, a no-skill pooled AUPRC reference of approximately 0.01798. Table 18 reports split and pair-distribution summaries; the accompanying CSV records each of the 210 sampled pair appearances. These are sampled prevalences, not a new full-valid-pixel census of all 200 pairs.

Readouts and calibration. The strict endpoint uses final flow, magnitude, image position, local flow roughness, and final internal norms. Endpoint-overmatch is a nonlinear MLP with 128- and 64-unit ReLU hidden layers and sigmoid calibration. Training and calibration use balanced sampled labels; this does not establish probability calibration at the natural test prevalence. We use ranking and retained-error comparisons, not a safety-calibration claim. The variance baseline (Durasov et al., 2024) combines flow variance and squared update energy. Aggregate convergence is the predefined primary readout; endpoint displacement is a feature-group ablation, not a post-hoc replacement primary model. Features include path length, endpoint-relative displacement, update moments, early/late ratios, adjacent-update cosines, and internal norms. The aggregate readout retains order-derived statistics. The shuffled control keeps the original aggregates and independently permutes feature streams within each example, perturbing alignment as well as chronology. It is not a chronology-only intervention.

Aggregation and uncertainty. AUPRC is computed by `average_precision_score`, not trapezoidal integration. Pooled metrics rank all held-out pixels within a seed and average the three seed metrics. Within-image differences average paired metrics over 207 eligible image-seed appearances; three single-class appearances are excluded from both within-image AUROC and AUPRC, but retained for pooled metrics. This estimand is appearance-weighted, not an equal-weighted mean over unique physical pairs.

There are 150 distinct physical pairs across 210 appearances: 99 occur once, 42 twice, and 9 three times. Each of 600 bootstrap draws resamples physical pairs with replacement and preserves all their appearances across seeds. The 95% percentile intervals are conditional on the fitted probes, not full training-run or checkpoint-generalization uncertainty. They are individual comparison intervals, without simultaneous multiplicity correction. An interval containing zero is unresolved, not proof of equivalence. Tables 15, 16, and 17 separate the ranking and retained-error estimands.

Table 15: **Corrected paired ranking comparisons.** Point differences and 95% physical-pair-clustered intervals (600 draws), conditional on fitted probes. Positive differences favor the first readout.

First minus second	Pooled AUROC	Within-image AUROC
Aggregate − MLP	$+0.0096$ $[-0.0015, +0.0183]$	-0.0117 $[-0.0266, +0.0027]$
Ordered − MLP	$+0.0037$ $[-0.0077, +0.0132]$	-0.0231 $[-0.0399, -0.0078]$
Displacement − MLP	$+0.0076$ $[-0.0024, +0.0170]$	-0.0168 $[-0.0304, -0.0043]$
Aggregate − Variance	$+0.0220$ $[+0.0131, +0.0307]$	$+0.0175$ $[+0.0053, +0.0295]$
Displacement − Variance	$+0.0201$ $[+0.0149, +0.0247]$	$+0.0124$ $[+0.0015, +0.0222]$
Variance − MLP	-0.0124 $[-0.0232, -0.0031]$	-0.0292 $[-0.0400, -0.0178]$
Aggregate − Endpoint	$+0.0455$ $[+0.0329, +0.0583]$	$+0.0351$ $[+0.0210, +0.0513]$
Last update − Endpoint	$+0.0105$ $[+0.0069, +0.0140]$	$+0.0099$ $[+0.0068, +0.0133]$
Ordered − Shuffled	$+0.0064$ $[+0.0032, +0.0096]$	$+0.0076$ $[+0.0026, +0.0128]$
Aggregate − Shuffled	$+0.0122$ $[+0.0083, +0.0163]$	$+0.0190$ $[+0.0095, +0.0283]$

Table 16: **Corrected paired ranking comparisons.** Point differences and 95% physical-pair-clustered intervals (600 draws), conditional on fitted probes. Positive differences favor the first readout.

First minus second	Pooled AUPRC	Within-image AUPRC
Aggregate − MLP	$+0.1015$ $[+0.0605, +0.1369]$	$+0.0582$ $[+0.0333, +0.0880]$
Ordered − MLP	$+0.0889$ $[+0.0469, +0.1262]$	$+0.0573$ $[+0.0316, +0.0873]$
Displacement − MLP	$+0.0484$ $[+0.0131, +0.0812]$	$+0.0405$ $[+0.0172, +0.0665]$
Aggregate − Variance	$+0.1727$ $[+0.1432, +0.1961]$	$+0.1390$ $[+0.1206, +0.1578]$
Displacement − Variance	$+0.1196$ $[+0.0974, +0.1389]$	$+0.1213$ $[+0.1050, +0.1377]$
Variance − MLP	-0.0712 $[-0.0932, -0.0450]$	-0.0808 $[-0.1033, -0.0524]$
Aggregate − Endpoint	$+0.1744$ $[+0.1245, +0.2171]$	$+0.0893$ $[+0.0639, +0.1161]$
Last update − Endpoint	$+0.0482$ $[+0.0341, +0.0598]$	$+0.0209$ $[+0.0133, +0.0292]$
Ordered − Shuffled	$+0.0167$ $[+0.0100, +0.0230]$	$+0.0093$ $[+0.0019, +0.0173]$
Aggregate − Shuffled	$+0.0294$ $[+0.0217, +0.0356]$	$+0.0102$ $[+0.0015, +0.0185]$

Table 17: **Corrected paired retained-error comparisons.** Point differences and 95% physical-pair-clustered intervals (600 draws), conditional on fitted probes. Negative differences favor the first readout; EPE is in pixels.

First minus second	Retained EPE at 80%
Aggregate − MLP	-0.0216 $[-0.0396, -0.0067]$
Ordered − MLP	-0.0200 $[-0.0389, -0.0039]$
Displacement − MLP	-0.0966 $[-0.1159, -0.0794]$
Aggregate − Variance	$+0.0701$ $[+0.0587, +0.0808]$
Displacement − Variance	-0.0049 $[-0.0118, +0.0006]$
Variance − MLP	-0.0916 $[-0.1096, -0.0756]$
Aggregate − Endpoint	-0.0234 $[-0.0407, -0.0089]$
Last update − Endpoint	-0.0046 $[-0.0102, -0.0002]$
Ordered − Shuffled	$+0.0081$ $[+0.0023, +0.0125]$
Aggregate − Shuffled	$+0.0066$ $[+0.0014, +0.0131]$

Table 18: **Corrected sampled prevalence.** Each split has 70 held-out pairs and 4,096 uniformly sampled valid pixels per pair. EPE is unrejected. Pair rates are sampled, not full-image censuses.

Seed	Overall (%)	Pair median (%)	Pair IQR (%)	Pair range (%)	EPE (px)
0	1.9228	1.0620	0.3296–2.3010	0.0488–10.4736	0.7256
1	1.7623	1.0986	0.2014–2.3132	0.0000–11.6455	0.7131
2	1.7090	0.9644	0.3235–2.0752	0.0000–13.4277	0.6678

Table 19: **Selective retention.** Mean retained EPE (pixels) at each nominal coverage, using pooled thresholds per seed. AURC integrates outlier fraction, not EPE, over coverage; these AURC values are descriptive without paired intervals. Lower is better for both.

Readout	100%	95%	90%	80%	AURC
Endpoint-overmatch MLP	0.7021	0.5648	0.4913	0.3988	0.002206
Aggregate convergence	0.7021	0.4890	0.4314	0.3772	0.002191
Variance + update energy	0.7021	0.4887	0.4044	0.3072	0.002665
Endpoint-displacement features	0.7021	0.4671	0.3874	0.3022	0.002250

Selective retention and scope. Coverage thresholds pool scores across images within each seed; pixels below the score quantile are retained. Score ties can make realized coverage differ slightly from nominal coverage. Table 19 expands the 80% result to 90%, 95%, and 100%; only the 80% paired EPE comparisons have the reported bootstrap intervals. Retention does not change the

underlying flow. The aggregate readout gains AUPRC over the MLP under both aggregations, but its AUROC advantage is unresolved. Variance has lower retained EPE than aggregate history. Endpoint displacement's retained-EPE difference from variance is unresolved. Ordered features improve ranking over shuffled features while worsening retained EPE. There is no uniform ranking of all readouts across these metrics.

Verification boundary. The correction materials contain the three saved prediction arrays, hashes, split reconstruction, pair-cluster draws, and independently recomputed metrics. A local CPU float32 parity check with the public KITTI checkpoint reproduces official RAFT's upsampled flows exactly at all 24 iterations for two synthetic RGB pairs, including a non-divisible image size requiring padding. This checks the corrected loop, not CUDA mixed precision, the file loader, or the actual remote cache. These checks do not by themselves reconstruct the complete remote trajectory cache or verify which image-reader backend produced every cached run. The original normalization-dependent sanity experiments have not been re-established on corrected caches. Other RAFT/Sintel caches require separate provenance checks; corrected D24 results do not certify them. No new SPair disjoint-fitting or WikiText training run is claimed in this revision.

A.13 Non-KITTI-Trained Checkpoints and Selective Utility

Question, selection, and population. These fixed-checkpoint follow-ups ask whether history adds reliability information beyond learned combined controls. They do not overwrite the earlier corrected experiment in Appendix A.12: the splits, readouts, and evaluated predictions differ. RAFT-Things and RAFT-KITTI use 24 updates. SEA-RAFT-M uses the public TartanAir/Chairs/Things checkpoint `Tartan-C-T432x960-M`, before KITTI-containing training, with four official refinements and five flows including learned initialization. Both models are from the RAFT family; this is not broad architectural coverage.

All 200 KITTI-2015 training pairs use the same permutation (seed 20270924): 120 fit, 40 validation, and 40 test. Fit/validation/test readouts use 2,048/4,096/8,192 uniformly sampled valid pixels per pair, respectively; sampling is label-independent, not class-balanced. Each test has 327,680 pixels. Sampled outlier prevalence is 14.2090% for RAFT-Things, 11.7572% for SEA-RAFT, and 1.1465% for RAFT-KITTI. One RAFT-KITTI pair has no outliers; within-pair AP excludes that pair, while pooled metrics retain it. Both classes occur in every test pair for the other two checkpoints. Backbone training and reliability-probe training are distinct: the KITTI-finetuned backbone has seen these images, but all probes hold out the 40 test pairs. Pair disjointness is verified, scene disjointness is not.

Each feature set has exactly four tuning candidates: standardized logistic regression with $C = 0.1$ or 1, and 100-round histogram boosting with 15 or 31 leaves, L2 coefficient 1, and no early stopping. Selection maximizes validation AP, without validation refitting or test-set selection. This matches tuning opportunities, not dimensionality or effective capacity. No checkpoint, horizon, or feature family was selected using these test scores. The dataset was previously inspected, so these remain exploratory follow-ups.

Observed representations. E contains final flow (two coordinates), magnitude, normalized image coordinates, spatial roughness, and final hidden/motion/correlation norms (nine features). It is not the complete recurrent endpoint state. SEA-RAFT's U adds four native mixture logits/raw log scales, two clamped log scales, the official weighted-log-scale heatmap, and mixture mean Laplace scale. B adds final-update vector/norm, temporal flow standard-deviation norm, update energy, and mean/max update norm to E (16 features for RAFT) or $E + U$ (24 for SEA-RAFT). Raw variance alone is an unfitted scalar comparator, not a claimed reproduction of every convergence-uncertainty method.

For RAFT, P contains 23 update norms and 24 endpoint distances; G is the normalized upper triangle of the last-eight-update Gram matrix plus log energy; W is summed signed adjacent-update area. $H = P + G + W$, giving 101 features with B. SEA-RAFT uses four update norms plus five endpoint distances, four-update Gram plus log energy, and one signed-area feature: P, G, W have 9, 11, and 1 features. The history length follows the deployed model rather than test-set shape selection. No ordering effect is inferred merely because a feature has a timestep index.

The RAFT response R branches at step 20: add positive/negative one-image- pixel flow-coordinate perturbations along each axis, keep hidden state fixed, and execute four updates per branch. Singular-

value, normalized-trace, and symmetric-remainder summaries at horizons 1, 2, and 4 cost 16 extra update evaluations. C instead executes 16 ordinary continuation updates. All labels still describe the original 24-step output. These are restricted finite response measurements, not full Jacobians or asymptotic stability estimates.

Table 20: **All fixed-protocol readouts.** No test-set winner selection. P: convergence profiles; G: temporal Gram; W: signed area; R: active perturbation response; C: equal-update continuation. All counts and feature definitions are specified below. Raw scalar controls are unfitted and are not substitutes for the learned combined baselines.

Checkpoint	Readout	AP	EPE80 (px)	Failure recall
RAFT-Things	E	0.7702	1.5204	77.97%
RAFT-Things	B	0.8157	1.0004	83.69%
RAFT-Things	$B + H$	0.8345	0.9196	85.66%
RAFT-Things	$B + R$	0.8276	1.0062	84.81%
RAFT-Things	$B + H + R$	0.8395	0.9272	85.78%
RAFT-Things	$B + H + C$	0.8384	0.9054	85.87%
RAFT-Things	R	0.7398	1.3234	78.96%
RAFT-Things	Raw variance	0.6029	1.3943	73.54%
RAFT-KITTI	E	0.2286	0.2462	88.58%
RAFT-KITTI	B	0.3005	0.2379	92.23%
RAFT-KITTI	$B + H$	0.3098	0.2349	92.39%
RAFT-KITTI	$B + R$	0.3094	0.2356	92.36%
RAFT-KITTI	$B + H + R$	0.3134	0.2344	93.03%
RAFT-KITTI	$B + H + C$	0.3123	0.2325	92.49%
RAFT-KITTI	R	0.1995	0.2964	79.72%
RAFT-KITTI	Raw variance	0.0883	0.2382	83.63%
SEA-RAFT	E	0.6756	1.1597	76.63%
SEA-RAFT	$E + U$	0.7832	0.8854	85.33%
SEA-RAFT	B	0.7920	0.8429	86.79%
SEA-RAFT	$B + P$	0.7942	0.8282	86.90%
SEA-RAFT	$B + G + W$	0.7929	0.8253	86.64%
SEA-RAFT	$B + P + G$	0.7932	0.8467	86.75%
SEA-RAFT	$B + P + W$	0.7952	0.8385	86.87%
SEA-RAFT	$B + H$	0.7941	0.8361	86.80%
SEA-RAFT	Native heatmap	0.6747	1.1182	74.78%
SEA-RAFT	Native mean scale	0.5687	1.1681	69.93%
SEA-RAFT	Raw variance	0.4607	1.3659	67.95%

Primary failures and secondary utility. The RAFT study's primary AP comparison was response added beyond full passive history; it is unresolved on both checkpoints and against equal-update continuation. History beyond B was a prespecified secondary comparison. SEA-RAFT's primary test was full history beyond B and is unresolved; simple convergence beyond $E + U$ was prespecified secondary. Reporting the secondary utility does not convert either primary test into a success.

Table 21: **Prespecified comparisons, including failed primary tests.** First minus second, with 600 paired physical-frame bootstrap draws. Positive AP and negative EPE differences favor the first method. These are unadjusted exploratory intervals conditional on the fitted split.

Contrast	AP difference [95% CI]	EPE80 difference [95% CI]
Things: history / B (secondary)	$+0.0188$ $[+0.0092, +0.0326]$	-0.0808 $[-0.2262, -0.0125]$
KITTI: history / B (secondary)	$+0.0093$ $[-0.0148, +0.0325]$	-0.0029 $[-0.0074, +0.0020]$
SEA: B / $E + U$ (secondary)	$+0.0088$ $[+0.0029, +0.0146]$	-0.0425 $[-0.0890, -0.0046]$
SEA: history / B (primary)	$+0.0022$ $[-0.0005, +0.0054]$	-0.0069 $[-0.0258, +0.0330]$
Things: response / history (primary)	$+0.0050$ $[-0.0017, +0.0153]$	$+0.0076$ $[-0.0491, +0.1045]$
KITTI: response / history (primary)	$+0.0036$ $[-0.0037, +0.0101]$	-0.0005 $[-0.0033, +0.0021]$

For RAFT-Things, within-pair AP also improves with history: $+0.0192$, CI $[0.0064, 0.0337]$. For SEA-RAFT, simple convergence's within-pair AP difference remains unresolved: $+0.0136$, $[-0.0016, 0.0295]$. Global ranking can use between-image difficulty; it is a different question from ranking pixels within each image, and neither aggregation substitutes for the other.

The subsequent RAFT component-removal analysis is explicitly post-hoc. Profiles alone give AP 0.8355 and EPE80 0.9147 on RAFT-Things; Gram plus area without profiles gives 0.8298 and 0.8984. Adding Gram plus area after profiles gives AP difference -0.0011 ($[-0.0038, 0.0007]$) and EPE difference $+0.0049$ ($[-0.0254, 0.0201]$). Thus geometry is not necessary to the observed useful history result. On SEA-RAFT, adding Gram after profiles is likewise unresolved on AP and EPE; its pooled-AUROC contrast is negative, -0.000664 ($[-0.001303, -0.000035]$), an unadjusted secondary result. Do not select the lowest table entry as a new winning descriptor.

Operating-budget accounting. All methods reject the globally highest-scoring 20% of sampled pixels; cutoff ties receive fractional expected retention. The independent reporting audit sorts low-to-high scores, recomputes the retained population, and reproduces every original EPE80/recall point and paired interval for all 27 readouts across the three checkpoints. It also verifies AP/AUROC points, prediction/model/source hashes, validation selections, and physical-pair membership. The 600 bootstrap draws resample the 40 physical pairs, retaining all their sampled pixels and using the same draw across readouts. Intervals condition on the fitted split and are not multiplicity-adjusted.

At this budget, history reduces RAFT-Things missed outliers by 12.05% ($[6.08, 17.89]\%$), equivalent to 27.92 fewer missed outliers per 10,000 evaluated pixels ($[10.13, 48.93]$). Simple convergence on SEA-RAFT reduces misses by 9.98% ($[4.05, 17.95]\%$), or 17.21 fewer per 10,000 ($[8.30, 27.74]$). These are post-hoc reporting transformations of the already specified recall metric, not new primary endpoints or new data. The relative retained-EPE reductions have intervals $[1.65, 14.14]\%$ and $[0.35, 9.83]\%$, respectively.

Table 22: **Operating-budget interpretation.** Post-hoc reporting transformations of the already specified 20% rejection metric, using the same saved predictions and 600 draws; not new primary endpoints. A positive relative reduction or negative count difference favors the first method. Counts are expected missed outliers per 10,000 evaluated pixels, with fractional cutoff ties.

Contrast	Missed reduction (%) [CI]	Count difference [CI]
Things: history / B (secondary)	$12.05\ [6.08, 17.89]$	$-27.92\ [-48.93, -10.13]$
KITTI: history / B (secondary)	$2.05\ [-9.26, 14.39]$	$-0.18\ [-1.62, +0.70]$
SEA: B / $E + U$ (secondary)	$9.98\ [4.05, 17.95]$	$-17.21\ [-27.74, -8.30]$
SEA: history / B (primary)	$0.06\ [-3.26, 3.07]$	$-0.09\ [-4.43, +5.34]$
Things: response / history (primary)	$0.85\ [-4.02, 8.12]$	$-1.74\ [-16.39, +9.00]$
KITTI: response / history (primary)	$8.39\ [-3.29, 12.03]$	$-0.73\ [-1.13, +0.22]$

Engineering checks and limits. The new caches use official forward, including RGB normalization and upsampling. KITTI uint16 readers match the official readers on every pair. Internal fields are aligned in padded coordinates before unpadding; RAFT perturbations convert image pixels to its one-eighth-resolution grid. GPU parity on pairs 0, 99, and 199 has zero maximum discrepancy for the checked flow histories; zero-perturbation continuation also matches. SEA-RAFT native outputs and heatmap match the official evaluation helpers. All-pair source/input/cache fingerprints and saved readout predictions were checked. These tests do not retroactively certify the older cache in Appendix A.12, whose provenance boundary remains stated.

Passive features require no extra backbone updates, but feature extraction and learned scoring are not free. Cached-feature prediction timings are not isolated end-to-end deployment measurements. A global budgeted ranking is not a calibrated online probability or fixed-threshold safety guarantee. The full-coverage flow predictions do not improve, and all three studies share the same inspected dataset and test pairs. We therefore claim bounded selective reliability utility, not a new SOTA flow method or universal trajectory benefit.

A.14 SPAIR-71K DIFFUSION-FEATURE AUDIT

Setup. We evaluate 12,234 SPair-71k test image pairs comprising 481 unique images, using nearest-neighbor PCK@0.1. Stable Diffusion inversion features at 11 saved timesteps use the Diffusion Hyperfeatures learned per-layer bottleneck projections. Ridge residualization fits unlabeled descriptors from the same evaluation images, and the orientation audit's PCA also fits those images. This is a transductive protocol, not the train-only protocol of the controlled residual audits. The best timestep is selected using evaluation PCK and is an oracle reference, not a validation-selected deployable choice. We retain the original descriptive measurements; no image-disjoint refit is claimed here. Aggregate PCK pools keypoints, whereas paired gaps average the per-pair PCK differences. Their weighting differs, so the paired gaps need not equal differences between the displayed aggregate PCK values.

Table 23: **Transductive SPair audit using the final timestep.** Timestep-aligned concatenation is competitive with the oracle best timestep. Residualizing against the weak final descriptor is a weaker conditioning test. Time-shuffled descriptors use independent image permutations and also break cross-image alignment.

Readout	Endpoint used for audit	$\mathrm{PCK}_{\mathrm{bbox}}$	$\mathrm{PCK}_{\mathrm{img}}$
Final saved timestep	final timestep	0.0582	0.0984
Best single timestep	timestep 2	0.6354	0.7029
Unordered mean over 11 timesteps	final timestep	0.5656	0.6369
Ordered 11-timestep trajectory	final timestep	0.6356	0.7013
Deterministic time-shuffled trajectory	final timestep	0.5283	0.5884
Residual unordered mean	final timestep	0.4942	0.5578
Residual ordered trajectory	final timestep	0.5645	0.6261
Residual deterministic time-shuffle	final timestep	0.4943	0.5498

Table 24: **Transductive SPair audit conditioning on the oracle best timestep.** Raw concatenation has no meaningful gain over that timestep. Residual concatenation exceeds mean pooling and independently shuffled descriptors; this is an alignment-sensitive comparison, not a chronology-only control.

Readout	$\mathrm{PCK}_{\mathrm{bbox}}$	$\mathrm{PCK}_{\mathrm{img}}$	Paired gap
Endpoint = best single timestep 2	0.6354	0.7029	–
Ordered 11-timestep trajectory	0.6356	0.7013	-0.0012 vs. endpoint
Residual unordered mean	0.1434	0.1780	–
Residual ordered trajectory	0.1865	0.2248	$+0.0406$ vs. residual unordered
Residual deterministic time-shuffle	0.1117	0.1377	–
Residual ordered trajectory	0.1865	0.2248	$+0.0667$ vs. residual shuffle

The original pair-row bootstrap gives 95% intervals $[-0.0033, 0.0009]$, $[0.0375, 0.0438]$, and $[0.0635, 0.0699]$ for the three paired gaps. The 12,234 rows reuse 481 images; these intervals do not preserve shared-image dependence and must not be interpreted as independent new-image generalization uncertainty. We report them only to document the original analysis, not as corrected significance evidence.

Timestep-window sensitivity. Smaller windows all include timestep 2. Aligned residual concatenation exceeds independently shuffled descriptors throughout, but exceeds residual mean pooling only for all 11 timesteps. These are descriptive differences under the same transductive fitting and alignment-sensitive control.

Table 25: **SPair timestep-window sensitivity under best-single endpoint conditioning.** All rows use timestep 2 as the endpoint for residualization. Paired gaps average same-pair $\mathrm{PCK}_{\mathrm{bbox}}$ deltas over 12,234 image pairs.

Timesteps	Resid. ord.–unord.	Resid. ord.–shuffle	Raw ord.–endpoint
$\{1, 2, 3\}$	-0.0068	$+0.0265$	$+0.0010$
$\{0, 1, 2, 3, 4\}$	-0.0041	$+0.0433$	-0.0001
$\{0, \ldots, 6\}$	-0.0070	$+0.0376$	-0.0004
$\{0, \ldots, 10\}$	$+0.0406$	$+0.0667$	-0.0012

Interpretation. This audit is not a SOTA descriptor or an inductive endpoint-insufficiency result. Its narrow observation is timestep-aligned residual correspondence after fitting on unlabeled evaluation features. If P applies the same timestep-block permutation to both descriptors, $(Pa)^\top(Pb) = a^\top b$; their normalized dot-product matches are unchanged. The implemented independent image permutations instead disrupt which timesteps align. Mean pooling also changes the representation, not only order. Consequently these controls do not isolate chronological adjacency or orientation. The nearly tied raw PCK suggests the best timestep already captures most utility under this task metric, while the residual finding motivates a disjoint-fitting follow-up.

SPair orientation-aware residual audit. PCA fits unlabeled residual deltas from the evaluation images, so this orientation extension is also transductive. Residual concatenation gives 0.1865 $\mathrm{PCK}_{\mathrm{bbox}}$, mean pooling 0.1434, and independent-image shuffling 0.1117. Local wedge gives 0.1006 versus 0.1005 under step permutation; Lévy area gives 0.1014 versus 0.0906, still below mean pooling. Adding local wedge to residual concatenation changes the mean paired PCK by -0.0319. The original pair-row intervals do not account for shared-image dependence. These measurements do not establish a useful chirality mechanism or an isolated chronology effect.

A.15 WikiText-103 Looped-Transformer Control

Setup. We train on WikiText-103's full training split with a 50,000-token vocabulary, 128-token contexts, $d = 192$, six tied inner loops, and 3,500 optimizer steps. All modes allocate 10,291,584 parameters and use the same six-loop architecture, but each backbone/readout is optimized end-to-end in a separate training run. Plain endpoint bypasses the fusion head: 221,760 parameters have no forward effect or gradient, leaving 10,069,824 active parameters. Endpoint-repeat and trajectory execute the same fusion head, with all 10,291,584 parameters active. Thus equal allocated counts do not establish equal active capacity or total readout compute. Endpoint-repeat, which feeds repeated final states to that head, is the executed-head control. Plain endpoint remains an important practical baseline and is reported separately. This is an end-to-end readout comparison, not language-modeling SOTA.

Table 26: **WikiText-103 looped-Transformer control.** Mean $\pm$ standard deviation over 10 seeds. Each configuration is trained end-to-end separately. History fuses its own backbone's loop states before the vocabulary head, not states shared with the endpoint-trained model.

Readout	Validation ppl.	Test ppl.
Endpoint-only	146.14 ± 1.83	135.97 ± 1.51
Endpoint-repeat	148.47 ± 1.17	138.44 ± 0.80
Endpoint+trajectory	136.66 ± 1.72	127.56 ± 1.46

Interpretation. The endpoint+trajectory configuration reduces mean test perplexity by 8.41 relative to endpoint-only and 10.88 relative to endpoint-repeat. Paired seed deltas are -8.41 with descriptive normal-approximation 95% interval [-9.81,-7.00] and -10.88 with interval [-12.11,-9.64]. Endpoint-repeat executes the same fusion-head architecture, controlling for simply allocating a larger head. All configurations use six loops, but training can produce different states and optimization behavior. Thus this is an end-to-end architecture/readout comparison, not a frozen-state isolation of information loss or a language-modeling SOTA claim.

Larger-scale stress run. The larger six-loop run uses $d = 384$, approximately 21.91M allocated parameters, and 7,000 optimizer steps over 3 seeds. Plain endpoint again bypasses the fusion head; only endpoint-repeat executes the same head as trajectory. The endpoint-versus-trajectory difference is unresolved, while endpoint-repeat performs worse. Width and training steps both increase, so this is a width/training-budget boundary, not an isolated capacity intervention.

Table 27: **Scaled WikiText-103 stress check.** Mean $\pm$ standard deviation over 3 seeds. The trajectory-minus-endpoint paired delta is $+0.10$ PPL with normal-approximation 95% interval $[-1.37, 1.57]$; trajectory-minus-endpoint-repeat is -4.55 with interval $[-5.29, -3.80]$.

Readout	Validation ppl.	Test ppl.
Endpoint-only	87.07 ± 0.85	82.23 ± 1.28
Endpoint-repeat	92.10 ± 0.96	86.87 ± 0.70
Endpoint+trajectory	87.04 ± 0.53	82.33 ± 0.05

This stress check does not support a scale-invariant trajectory advantage. The comparison with endpoint-repeat remains informative, but the larger plain endpoint performs comparably under this audit. Three seeds and the descriptive normal-approximation intervals do not establish equivalence or show how information was absorbed during training.

Loop-count sensitivity. We also repeat the matched-budget comparison at 4 and 8 tied loops using 3 seeds each. The endpoint+trajectory advantage persists at both loop counts, while absolute perplexity is not monotonic in loop count because the models are not retuned for each depth.

Table 28: **WikiText-103 loop-count sensitivity.** Test perplexity for matched-budget readouts. The paired trajectory-minus-endpoint deltas are -6.06 for 4 loops, -8.41 for 6 loops, and -11.25 for 8 loops.

Loops	Seeds	Endpoint	Endpoint-repeat	Endpoint+trajectory
4	3	131.67 ± 1.61	133.17 ± 2.82	125.61 ± 0.51
6	10	135.97 ± 1.51	138.44 ± 0.80	127.56 ± 1.46
8	3	140.30 ± 1.58	140.86 ± 1.83	129.06 ± 1.04

Residualized hidden-state audit. Finally, we run the endpoint-residualized audit on the hidden states of 3 independently trained 6-loop trajectory models. For each seed, we collect 200k train positions and 50k validation/test positions, fit train-only ridge maps $h_t \approx A_t h_6 + b_t$ for $t < 6$, and train residual next-token probes on $(\tilde{h}_1, \ldots, \tilde{h}_5)$. Table 29 summarizes the endpoint predictivity and residual readout results. Even after conditioning on h_6, the ordered residual probe beats the permutation-augmented residual control. The gap is small but consistent, so we treat it as supportive evidence for distributed residual order in loop-time states beyond a linearly endpoint-predictable history. This residual-only probe is not a competitive language model and is not the end-to-end consequence evidence; that role belongs to the matched readout and loop-count comparisons above.

Table 29: WikiText-103 residualized hidden-state audit. The residual-only probe is a mechanism check, not a competitive language-model result.

Quantity	Value
Mean $R^2(h_t \mid h_6)$ for $t = 1, \ldots, 5$	$0.8020,\ 0.8941,\ 0.9500,\ 0.9818,\ 0.9962$
Ordered residual probe, test perplexity	4477.09 ± 154.16
Permutation-augmented residual probe, test perplexity	4703.68 ± 182.69
Ordered residual probe, test accuracy	0.1211 ± 0.0011
Permutation-augmented residual probe, test accuracy	0.1153 ± 0.0009
Ordered $-$ permuted residual NLL delta	-0.0493 ± 0.0047
Ordered $-$ permuted residual accuracy delta	$+0.0058 \pm 0.0010$

Residual orientation audit. To test whether trained loop dynamics contain orientation-sensitive order signal, we rerun the residual audit on the 6-loop trajectory models and compute local-wedge and Lévy-area descriptors after ridge-conditioning earlier loop states on h_6. The projection is fit on train residual displacements only, uses 16 dimensions, and yields 120 antisymmetric coordinates. These residual-only probes are not competitive language models; the question is whether ordered orientation features beat step-permuted and time-reversed orientation controls under the same next-token probe.

Table 30: **WikiText-103 residual orientation audit.** Mean $\pm$ standard deviation over 3 seeds. Local wedge and Lévy-area descriptors beat step-permuted controls by test-NLL deltas -0.7075 ± 0.0202 and -0.9312 ± 0.0447, respectively. Forward probes beat time-reversed inputs by NLL margins 1.8648 ± 0.0170 and 2.7290 ± 0.1386.

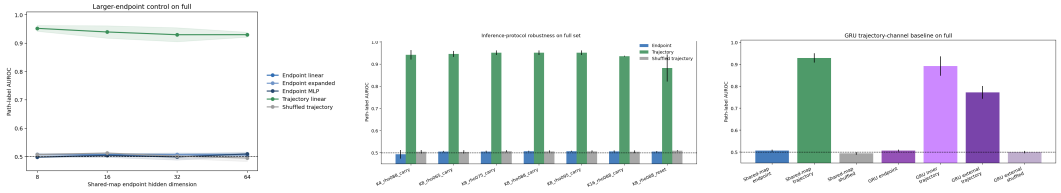
Feature	Test NLL	Test accuracy
Local wedge	8.8714 ± 0.0238	0.0776 ± 0.0043
Local wedge, step-permuted	9.5789 ± 0.0036	0.0192 ± 0.0007
Local wedge, time-reversed	10.7362 ± 0.0099	0.0063 ± 0.0011
Lévy area	8.6371 ± 0.0495	0.0778 ± 0.0044
Lévy area, step-permuted	9.5683 ± 0.0048	0.0222 ± 0.0020
Lévy area, time-reversed	11.3661 ± 0.0892	0.0050 ± 0.0006

A.16 Additional Control Details

Endpoint capacity. Increasing endpoint dimension from the reported budget and strengthening endpoint probes does not recover the path label: endpoint-only stays near 0.51, while trajectory remains above 0.92.

Inference protocol. Varying rollout length, contraction clip, and carry-vs-reset initialization preserves the separation: endpoint and shuffled controls remain near chance, while trajectory AUROC ranges from 0.8828 ± 0.0609 to 0.9522 ± 0.0103.

Recurrent trajectory-access control. Giving the GRU baseline explicit trajectory access substantially improves it, supporting the broader endpoint-plus-trajectory claim. GRU inner trajectory reaches 0.8917 ± 0.0442 and external trajectory reaches 0.7717 ± 0.0292, compared with 0.5069 ± 0.0044 for GRU endpoint. The shared-map trajectory remains strongest at 0.9287 ± 0.0212.



(a) Endpoint capacity (b) Inference protocol (c) Recurrent trajectory access

Figure 7: **Additional controls.** The trajectory effect is not explained by endpoint capacity, inference protocol, or denying recurrent baselines access to trajectory channels.