RESIDUAL GEOMETRY SEPARATES STATE MOTION FROM PREDICTIVE CHANGE IN WEIGHT-TIED ITERATIVE REFINEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

For weight-tied networks, the relative hidden-state residual is commonly applied as a convergence diagnostic, yet it can remain large when predictions have frozen and can look similar while predictions are still changing. We explain this mismatch by decomposing a raw update into shared scale, unit-wise scale heterogeneity, and within-unit direction. The decomposition is exact and exposes which motion a readout can see. Across nine 3D refiners, the terminal residual clusters near 0.34 and consists almost entirely of shared scaling, with no label flips. The fully retained $T = 96$ double-precision bundles attribute the late directional plateau in the `float32` runs to a storage floor. An extrapolating prefix-sums model and three pretrained LoopFormer checkpoints instead remain orientation- and output-active. Equal-energy interventions confirm that component energy and decision sensitivity are different: direction flips 6–13% of 3D points and 63–75% of LoopFormer tokens, while shared scale flips none. A projective constraint makes the realised state contract but does not stabilise prediction. These results motivate a geometry-and-readout halting gate that requires resolved orientation and output change to settle and refuses unsafe truncation of an unfinished internal clock. On held-out LoopFormer sequences, that refusal is necessary: matched-cost dynamic rules lose more than a nat while the native schedule loses only 0.03 nats at 25% lower compute.

1 INTRODUCTION

Weight tying turns depth into computation: a network repeatedly applies $H_{t+1} = F_\theta(H_t, x)$ and may be run for more or fewer iterations without changing its parameter count. A common operational shortcut is to monitor

$$r_t = \frac{\|H_{t+1} - H_t\|_F}{\|H_t\|_F}$$

and interpret a small value as convergence or a large value as unfinished computation. That interpretation assumes that all state motion matters to the prediction in the same way.

Our measurements show the opposite. A dense 3D refiner reaches $r_t \approx 0.34$ and stays, yet its labels freeze. A prefix-sums model at $16\times$ its training length has a smaller residual, 0.17, while its predictions and accuracy keep changing. Three LoopFormers have residuals between 0.72 and 1.00 and remain output-active at their native iteration budgets. The raw residual therefore groups together systems with different geometry, readouts, and task behaviour. The scalar residual obscures whether the state changes in scale or orientation and whether that motion affects predictions.

We recover that distinction in two stages. First, Equation (1) separates global norm change from orientation change. Second, an orthogonal unit-wise decomposition splits the update into shared scaling, heterogeneous scaling across units, and directional motion within each unit. The resulting account is simple: the late 3D state mostly grows along itself, whereas prefix-sums extrapolation and LoopFormer continue to rotate. Readout calibration then determines the functional consequence. A normalised readout removes positive shared scaling; a softmax may preserve the label while its confidence saturates; directional motion can continue to move decision boundaries.

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.

This view also changes what a usable stopping rule should measure. We propose a *geometry-and-readout gate*: halt only after resolved orientation change and output distribution change are both small for several steps, and only when stopping does not truncate an unfinished exogenous clock. The gate deliberately abstains when those conditions fail and falls back to the native schedule. The rule is intended as a diagnostic criterion.

Original contributions. Our contributions are:

1. We give an exact residual anatomy and its readout consequence, turning a raw fixed-point residual into three interpretable geometric components. The definitions stay in the main text; derivations, numerical resolution, and implementation details are in the appendix.
2. We verify the explanation across samples, datasets, models, and update rules: nine independently trained 3D refiners, two input regimes of a pretrained prefix-sums checkpoint, three pretrained LoopFormer checkpoints, a same-task contractive control, and three projectively trained seed pairs.
3. We convert the theory into an operational halting gate and test the failure mode it is designed to catch. Frozen residual, orientation, logit, and probability rules all lose badly to a native schedule at matched cost on time-conditioned LoopFormers; an equal-cost retiming control attributes most of that loss to leaving the internal clock unfinished.

Relation to prior work. Deep-equilibrium models explicitly seek a fixed point (Bai et al., 2019; 2021; Winston & Kolter, 2020; Revay et al., 2020); recurrent-depth models reuse parameters without that guarantee (Dehghani et al., 2019), and added depth can extrapolate or overthink (Bansal et al., 2022; Anil et al., 2022; Labovich, 2026; Movahedi et al., 2026; Kim et al., 2026; Zhang et al., 2026). Adaptive-computation methods learn halting signals (Graves, 2016; Banino et al., 2021; Figurnov et al., 2017; Bae et al., 2025; Moosa et al., 2026). We ask the complementary measurement question: before learning or thresholding a residual, what part of the state is moving, can the readout see it, and is the proposed stop numerically and temporally valid?

2 From raw residual to functional motion

2.1 Two exact decompositions

Let $\rho_t = \|H_{t+1}\|_F / \|H_t\|_F$ and let

$$q_t = \left\| \frac{H_{t+1}}{\|H_{t+1}\|_F} - \frac{H_t}{\|H_t\|_F} \right\|_F$$

be the stable normalised-state change. If ϕ_t is the angle between the vectorised states, then

$$r_t^2 = (\rho_t - 1)^2 + \rho_t q_t^2 = (\rho_t - 1)^2 + 2\rho_t(1 - \cos\phi_t). \tag{1}$$

Thus r_t can be large because the state changes length, direction, or both.

To expose spatial or token-wise structure, write the state as unit vectors $H_t = [h_{t,1}, \dots, h_{t,N}]^\top$, let $\Delta_i = h_{t+1,i} - h_{t,i}$, and define

$$a_i = \frac{\langle \Delta_i, h_{t,i} \rangle}{\|h_{t,i}\|^2}, \qquad \bar{a} = \frac{\sum_i a_i \|h_{t,i}\|^2}{\|H_t\|_F^2}. \tag{2}$$

The three components are

$$\Delta_i^{\text{shared}} = \bar{a} h_{t,i}, \quad \Delta_i^{\text{hetero}} = (a_i - \bar{a}) h_{t,i}, \quad \Delta_i^{\text{direction}} = \Delta_i - a_i h_{t,i}. \tag{3}$$

They are pairwise orthogonal, hence

$$r_{\text{raw}}^2 = r_{\text{shared}}^2 + r_{\text{hetero}}^2 + r_{\text{direction}}^2, \qquad r_k = \|\Delta^k\|_F / \|H_t\|_F. \tag{4}$$

Equation (3) is a definition; the orthogonality proof and the treatment of zero-norm units are in App. D.

The shared coefficient is $\bar{a} = \rho_t \cos\phi_t - 1$, not generally $\rho_t - 1$ from Equation (1); the latter measures endpoint norm change. Their difference is small only when the endpoints are nearly collinear. The two non-shared energies sum to $\rho_t^2 q_t^2 (1 - q_t^2/4)$, but separating heterogeneity from direction requires the unit-wise coefficients in Equation (2).

Readout consequence. For a differentiable readout g, positive scale invariance $g(cH) = g(H)$ implies $J_g(H)H = 0$ and therefore $J_g(H)\Delta^{\text{shared}} = 0$. Shared scale is locally invisible to such a readout, whereas no corresponding null follows for either non-shared component. Without scale invariance, shared motion may still change logits or confidence without changing the argmax. This separates three questions that raw r_t conflates: whether the state moves, whether the readout moves, and whether the task decision improves. The finite-perturbation qualification and proof are in App. E.

2.2 A GEOMETRY-AND-READOUT HALTING GATE

Let $p_t = g(H_t)$ be the predictive distribution, let m be a patience window, and let χ_t indicate that stopping at t either preserves completion of an exogenous time schedule or explicitly retimes the final step to complete it. We define the gate

$$\text{halt}(t) = 1 \iff \chi_t = 1 \text{ and } \max_{j=t-m+1}^{t} q_j \le \tau_q \text{ and } \max_{j=t-m+1}^{t} D_{\text{JS}}(p_{j+1}, p_j) \le \tau_p. \quad (5)$$

The thresholds must clear the measured numerical floor, be selected on held-out data, and be checked against task loss at matched compute. A failed condition means abstention and use of the native schedule. Raw r_t is intentionally absent because shared scale can keep it large after predictions settle. Conversely, small q_t is not enough if the readout or task is unstable. Equation (5) withholds a halting decision when orientation or output change remains above threshold or when the internal clock is unfinished.

3 EXPERIMENTAL PROTOCOL

Families and replication. The 3D system replaces COSeg's stack of distinct correlation aggregators (An et al., 2024) with one tied layer trained for eight iterations. We audit three confirmatory S3DIS seeds, one ScanNet replication, and five exploratory S3DIS configurations to $T = 64$, with 100 episodes per checkpoint. The pretrained depth-recurrent prefix-sums model (Bansal et al., 2022) is evaluated at its 32-bit training length and at 512 bits ($16\times$ extrapolation). Three pretrained 0.2–0.3B LoopFormer checkpoints (Jeddi et al., 2026) are matched at 24 block-equivalent evaluations and audited on 256 held-out sequences each. The training run is the replicate for 3D; prefix-sums conditions and LoopFormer sequences are repeated measurements within checkpoint.

Measurements and interventions. At each step we record r_t, ρ_t, q_t, the three component energies, centred-logit change, probability Jensen–Shannon divergence, label flips, and the task metric. We inject equal Frobenius energy along each component to separate energy share from readout sensitivity. A same-task contractive checkpoint is a positive control for the probe. Three projective runs rescale every update to the input norm and test whether forcing realised contraction stabilises predictions. All forward states are `float32`; diagnostic reductions use `float64`. A matched three-seed recurrence in `float64` distinguishes arithmetic from state-storage floors. Full recurrence, sampling, aggregation, precision, and provenance details are in Apps. A–C.

4 RESULTS

4.1 RAW RESIDUAL GIVES THE WRONG FUNCTIONAL GROUPING

Table 1 presents the central contradiction. The 3D refiner has the largest raw residual of the three conditions, yet q_T, probability change, and label flips are at their numerical floor. Prefix-sums extrapolation has a smaller r_T but places more than 99.9% of its residual energy in orientation; labels still change, terminal probability Jensen–Shannon divergence is 0.021, and accuracy rises from 0.579 to 0.661 after the training depth. At the trained input length, raw state, orientation, output, and task all settle together.

Figure 1 shows the same separation over depth. In 3D, r_t stays near 0.34 while q_t and label flips fall by five and four decades respectively. In prefix-sums extrapolation, orientation and output change persist. At a threshold of 10^{-2}, raw residual classifies both the prediction-stable 3D refiner and the output-active extrapolating prefix-sums model as *unconverged*. The already converged trained-length prefix-sums condition is excluded. Despite the difference between the raw values 0.34 and 0.17, the same coarse decision misses the distinction that predicts their behaviour.

Table 1: **The same raw residual can describe opposite functional regimes.** Terminal episode means. Norm share is $(\rho - 1)^2/r^2$; task change is measured from each model's training depth and is interpreted only within task.

System	r_T	$\lvert\rho_T - 1\rvert$	q_T	norm share	flip	task Δ
prefix-sums, trained length	7.8×10^{-8}	$\leq 6.0\times10^{-8}$	7.8×10^{-8}	$[0, .62]$	.000	+.0000
prefix-sums, 16× length	.1694	.0030	.1692	5.1×10^{-4}	.039	+.0816
3D refiner, seed 123	.3403	.3403	2.9×10^{-8}	$> 1 - 1.1\times10^{-14}$	.000	-.0065

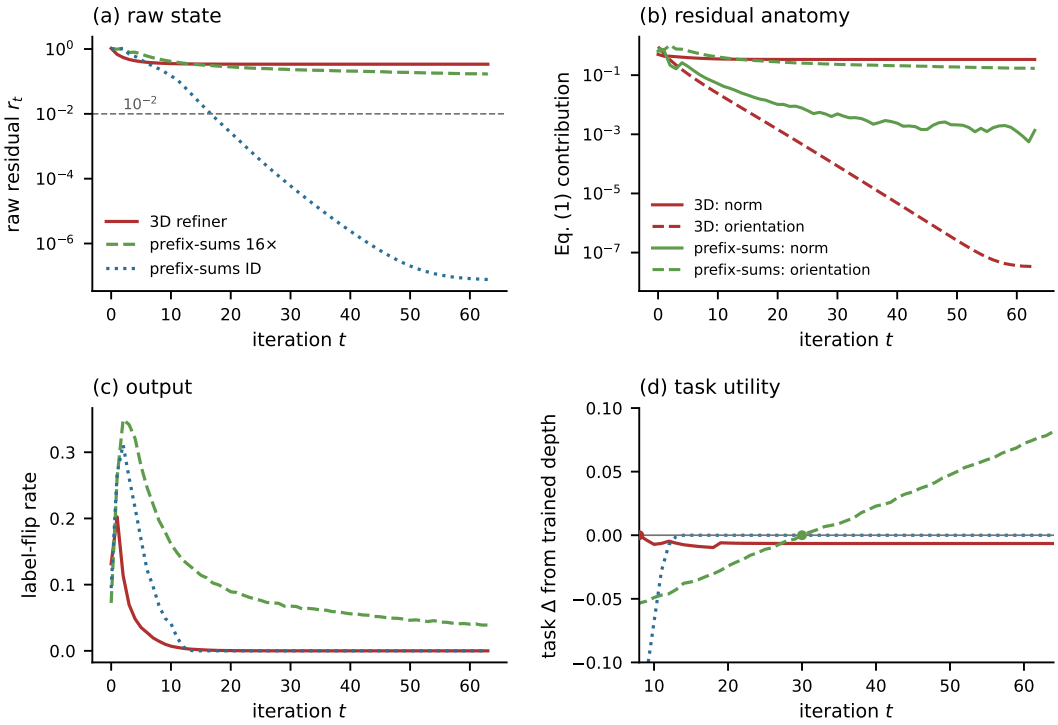

Figure 1: **Residual anatomy, output change, and task utility disagree with raw magnitude.** (a) Raw residual. (b) Norm and orientation terms in Equation (1). (c) Label changes. (d) Task change from training depth; task metrics are compared only within curve. The 3D residual stays large because of scale, while extrapolating prefix-sums remains direction- and output-active.

4.2 The 3D anomaly replicates and is not a precision artefact

Across nine independently trained 3D checkpoints, r_T and $\rho_T - 1$ agree to at most 5.96×10^{-8}, two `float32` ulp at this scale. The worst of 900 episodes has an orthogonal step share below 1.23×10^{-7}, so less than 1.6×10^{-14} of terminal residual energy is identified as non-shared. Table 2 shows the confirmatory seeds and dataset replication; the five configuration-robustness rows are in Table A1.

The observed terminal $q \approx 2.9 \times 10^{-8}$ is a storage limit rather than a measured zero. Over three seeds and 20 matched episodes each, a `float64` recurrent core lowers the late minimum by 4.917 dex (CI95 $[-4.922, -4.911]$), while double arithmetic with `float32` storage changes it by only 0.052 dex. The fully retained $T = 96$ bundle supports this claim. The separately frozen $T = 128$ summary is sensitivity evidence because two seeds' raw episode arrays are unavailable (App. C).

4.3 The decomposition explains which motion matters

Before the floor, the shared component carries 99.84–99.85% of the 3D update energy. Of the small remainder, direction exceeds unit-wise scale heterogeneity by 212–467× in the confirmatory seeds.

Table 2: **Large late residuals replicate as shared scale drift.** Each row contains 100 episodes. Gap is $|r_T - (\rho_T - 1)|$; non-shared is the worst episode's exact orthogonal share. The five exploratory configurations give the same regime (Table A1).

Run	data	r_T	ρ_T	gap	non-shared $\leq$	κ
seed 123	S3DIS	.340251	1.340251	3.0×10^{-8}	1.2×10^{-7}	.7500
seed 777	S3DIS	.342527	1.342527	0	1.2×10^{-7}	.7488
seed 2026	S3DIS	.342604	1.342604	3.0×10^{-8}	1.2×10^{-7}	.7476
ScanNet fold 1	ScanNet	.345405	1.345405	5.4×10^{-8}	1.1×10^{-7}	.7470

Table 3: **The small non-shared motion is directional, not unit-wise scaling.** Component energy shares are means over the numerically resolved window $t \in [8, 30]$. The same probe recovers order-one heterogeneity in a contractive control.

Run	E_{shared}	E_{hetero}	$E_{\text{direction}}$	hetero/non-shared	μ
seed 123	.998401	5.0×10^{-6}	1.6×10^{-3}	.0021	1.00224
seed 777	.998528	9.8×10^{-6}	1.5×10^{-3}	.0047	1.00262
seed 2026	.998466	7.9×10^{-6}	1.5×10^{-3}	.0034	1.00162
contractive control	.122585	.211	.666	.2418	.87659

The contractive positive control returns $E_{\text{hetero}} = 0.211$ and $E_{\text{direction}} = 0.666$, showing that the same instrument resolves both when present (Table 3).

Equal-energy injection supplies the functional test that energy shares cannot. At $t = 25$, shared scale flips no label in any confirmatory seed, while direction flips 6.0–13.4% of points. At $t = 8$, direction still flips 23–267$\times$ more labels even though shared injection moves centred logits four times farther. The late shared mode therefore reaches the logits but mostly rescales already saturated scores. It is neither numerically absent nor decision-active. Across depth, log flip rate versus $\log q_t$ has slope 0.84–1.04, while the same points are nearly vertical against raw r_t. At the terminal 3D step, centred-logit change remains 0.340 while Jensen–Shannon divergence and label flips are zero. This pattern is consistent with saturation rather than readout invariance. Full component values and on-manifold deletion controls are in App. G.

Operational reading. A large raw residual requires separating $|\rho_t - 1|$ from q_t and then inspecting the readout. Small r_t establishes only raw-state stationarity within the measured precision and horizon. Neither case alone establishes task convergence.

4.4 The explanation transfers across models and readouts

All three LoopFormers occupy the opposite regime from 3D: terminal q_T is 0.72–1.07, shared energy is 0.13–0.51, the last label flip occurs at the configured horizon, and terminal Jensen–Shannon divergence is 0.053–0.423. Equal-energy direction flips 63–75% of tokens, whereas positive shared scaling flips none (Table 4).

The shared null is explained by the RMS normalisation immediately before the language head. Injecting before it moves centred logits by only 4.0–4.5×10^{-7}. Injecting the same mode one module later moves them by exactly 0.5 and changes NLL by $+1.17$ to $+1.71$ nats, a ratio near 10^6. Thus the null belongs to the readout position rather than an information-free state (App. H). Per-token settling is not identified either: residuals end 8.7–15.7 times above their minima. On the held-out slice, the rule censors 44–45% of tokens. The resulting depth spread matches a token-identity permutation null, leaving token-specific settling depths unidentified.

4.5 Changing the algorithm confirms the limit of contraction

A projective update rescales every state to its pre-update norm. It makes the realised orientation factor contractive in all three retrained seeds ($\mu = 0.923$–0.991) and turns the update directional. Yet every projective run still changes labels at $t = 127$, and paired trained-depth validation differences

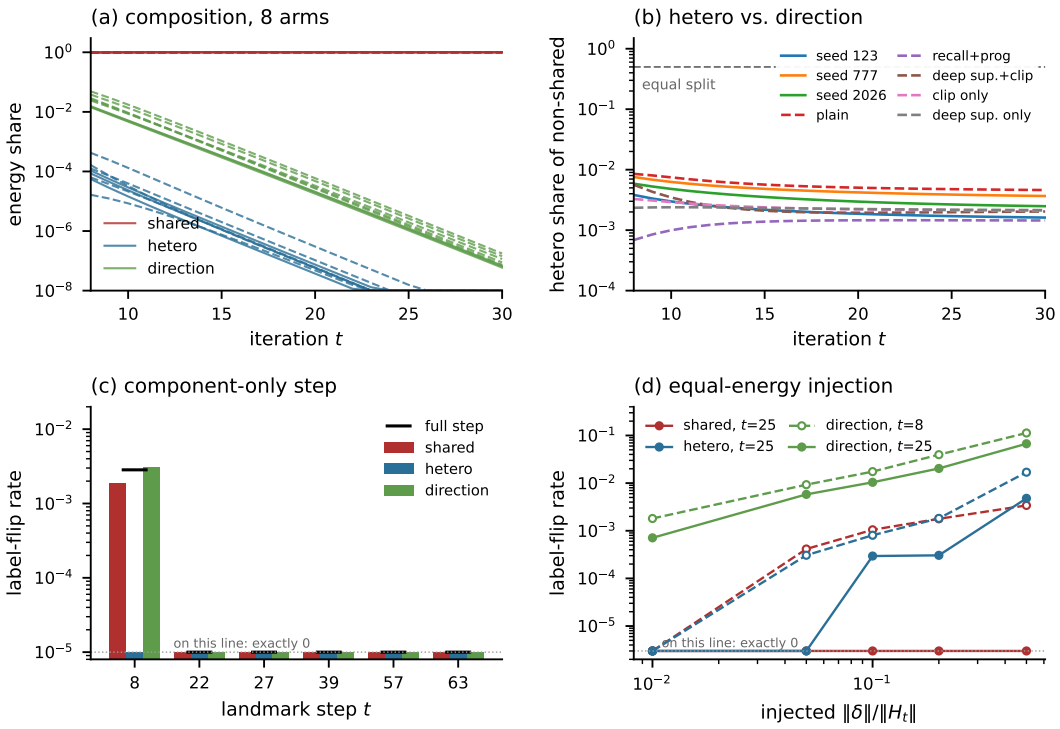

Figure 2: **Energy dominance and decision sensitivity are different.** (a) Three component shares in eight 3D runs. (b) Heterogeneous scale is a small part of the non-shared motion. (c) Component-only natural steps. (d) Equal-energy interventions: at $t = 25$, shared scale flips no label while direction flips about one in ten.

Table 4: **Three pretrained LoopFormer checkpoints remain orientation- and output-active.** Means over 256 held-out sequences per checkpoint at 24 block-equivalent evaluations. Equal-energy injections have amplitude $0.5\|H\|$.

Checkpoint	r_T	q_T	E_{shared}	last/iters	shared	direction
3 blocks × 8	.996	1.073	.513	8/8	.000	.751
2 blocks × 12	1.001	1.058	.424	12/12	.000	.744
1 block × 24	.724	.716	.126	24/24	.000	.634

straddle zero (Table 5). Contraction is therefore not sufficient for prediction stability, and removing the original scale mode is not itself a halting solution.

4.6 A USABLE GATE MUST BE ALLOWED TO ABSTAIN

The observed trajectories are consistent with the intended decisions of Equation (5). The matched-cost halting test covers the LoopFormer checkpoints. The gate is not blocked by the large 3D r_t once q_t and output change settle; it does not license stopping on prefix-sums extrapolation or LoopFormer while orientation and predictions remain active. For a model with an internal time coordinate, χ_t also rejects naive truncation.

We internally pre-specified tests of that failure mode with thresholds on q, raw r, Jensen–Shannon divergence, and centred-logit change. Thresholds were cost-matched on 256 validation sequences, frozen, and entered once on a disjoint 256-sequence slice. At 25% lower compute, the native schedule loses only 0.026–0.028 nats, while the best qualifying dynamic signal loses 1.67–2.73 nats (Table 6). Every paired interval excludes zero by more than a nat.

Table 5: **Making the realised state contract does not stabilise prediction.** The projective update preserves state norm by construction. Values use matched controls; validation $\Delta mIoU$ is projective minus control at training depth.

Seed	μ control	μ projective	projective r_T	$E_{\text{direction}}$	val. $\Delta mIoU$	last flip
123	1.00224	.99124	.05392	.9919	-.0089	127
777	1.00262	.94053	.00911	.9942	+.0293	127
2026	1.00162	.92310	.00728	.9856	-.0205	127

Table 6: **Naive dynamic triggers are unsafe on time-conditioned checkpoints.** Next-token NLL increase versus the full budget. Native and best-signal columns use the held-out 256-sequence slice; the retimed column is a validation-slice mechanism diagnostic. The best frozen dynamic signal is compared at matched cost with the model's native k-step schedule; lower is better.

Checkpoint	saving	native ΔNLL	best ΔNLL	best $-$ native, CI95	retimed ΔNLL
3 blocks $\times$ 8	25%	+.0264	+1.6720	$[+1.399, +1.892]$	+.0916
3 blocks $\times$ 8	50%	+.0947	+3.4906	$[+3.231, +3.552]$	+.2207
2 blocks $\times$ 12	25%	+.0275	+2.7345	$[+2.634, +2.773]$	+.1425
2 blocks $\times$ 12	50%	+.0835	+2.9391	$[+2.637, +3.058]$	+.3134
1 block $\times$ 24	25%	+.0278	+1.9451	$[+1.792, +2.040]$	+.4403
1 block $\times$ 24	50%	+.1252	+3.4138	$[+3.128, +3.456]$	+.6918

A validation-slice mechanism diagnostic retimes only the final step at the same number of steps. It recovers 85–97% of the truncation gap across checkpoints (median over budgets). The main failure is therefore an unfinished internal clock on these checkpoints. The practical rule is conservative: validate q and output thresholds above their floors, require a clock-safe stop, and retain the native schedule whenever the gate abstains.

5 Limitations and conclusion

Only the 3D family has multiple independent training runs. Prefix-sums and LoopFormer claims are checkpoint-conditional. Equal-energy injections measure local finite sensitivity, and most 3D shared-component deletions leave the readout's operating range. The projective experiment has three runs per condition and changes learned dynamics as well as norm growth. LoopFormer sequences come from a disjoint training-shard slice rather than an official validation set. No finite horizon proves that a later fixed point does not exist. Expanded limitations are in App. L.

The main empirical fact is that raw residual magnitude does not determine predictive convergence. Equation (1) and the unit-wise split show why: shared scaling explains the large but decision-inactive 3D residual, while orientation explains continued change in prefix-sums and LoopFormer. Readout interventions, positive controls, projective training, precision tests, and matched-cost halting all support the same boundary. A practical stopping procedure should therefore monitor resolved orientation and output change, respect the model's time parameterisation, and abstain when those conditions do not justify a stop. This separates raw-state stationarity, output convergence, and task utility rather than treating any one of them as a proxy for the other two.

AI use statement

OpenAI ChatGPT and Anthropic Claude Opus models assisted with targeted language polishing and readability edits during manuscript revision. The author reviewed the resulting wording and remains responsible for the final text and claims.

Reproducibility statement

Section 3 specifies the recurrence, metrics, aggregation, and replication units. Appendices A–Q document data splits, numerical resolution, controls, provenance, environments, and implementation

checks. The retrieval boundary of the secondary $T = 128$ sensitivity extension is disclosed in App. C.

ACKNOWLEDGMENTS

GPT and Claude Opus models assisted with targeted language polishing during manuscript revision. The author reviewed the resulting wording and remains responsible for the final text and claims.

REFERENCES

Zhaochong An, Guolei Sun, Yun Liu, Fayao Liu, Zongwei Wu, Dan Wang, Luc Van Gool, and Serge Belongie. Rethinking few-shot 3d point cloud semantic segmentation. In *CVPR*, 2024.

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 *NeurIPS*, 2022. arXiv:2211.09961.

Sanjeev Arora, Zhiyuan Li, and Kaifeng Lyu. Theoretical analysis of auto rate-tuning by batch normalization. In *ICLR*, 2019.

Sangmin Bae, Yujin Kim, Reza Bayat, Sungnyun Kim, Jiyoun Ha, Tal Schuster, Adam Fisch, Hrayr Harutyunyan, Ziwei Ji, Aaron Courville, and Se-Young Yun. Mixture-of-recursions: Learning dynamic recursive depths for adaptive token-level computation. *arXiv preprint arXiv:2507.10524*, 2025.

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. Deep equilibrium models. In *NeurIPS*, 2019.

Shaojie Bai, Vladlen Koltun, and J. Zico Kolter. Stabilizing equilibrium models by jacobian regularization. In *ICML*, 2021. arXiv:2106.14342.

Andrea Banino, Jan Balaguer, and Charles Blundell. Pondernet: Learning to ponder. *arXiv preprint arXiv:2107.05407*, 2021.

Arpit Bansal, Avi Schwarzschild, Eitan Borgnia, Zeyad Emam, Furong Huang, Micah Goldblum, and Tom Goldstein. End-to-end algorithm synthesis with recurrent networks: Logical extrapolation without overthinking. In *NeurIPS*, 2022. arXiv:2202.05826.

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Łukasz Kaiser. Universal transformers. In *ICLR*, 2019.

Michael Figurnov, Maxwell D. Collins, Yukun Zhu, Li Zhang, Jonathan Huang, Dmitry Vetrov, and Ruslan Salakhutdinov. Spatially adaptive computation time for residual networks. In *CVPR*, 2017.

Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, 4th edition, 2013.

Alex Graves. Adaptive computation time for recurrent neural networks. *arXiv preprint arXiv:1603.08983*, 2016.

Ahmadreza Jeddi, Marco Ciccone, and Babak Taati. LoopFormer: Elastic-depth looped transformers for latent reasoning via shortcut modulation. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=RzYXb5YWBs.

Bum Jun Kim, Kohei Hayashi, Shunsuke Kamiya, Masanori Koyama, Yusuke Iwasawa, and Yutaka Matsuo. Looped transformers with source-centered state evolution. *arXiv preprint arXiv:2607.27656*, 2026.

Asher Labovich. Stability and generalization in looped transformers. *arXiv preprint arXiv:2604.15259*, 2026.

Ibraheem Muhammad Moosa, Suhas Lohit, Ye Wang, Moitreya Chatterjee, and Wenpeng Yin. Understanding dynamic compute allocation in recurrent transformers. *arXiv preprint arXiv:2602.08864*, 2026.

Sajad Movahedi, Vera Milovanović, Shlomo Libo Feigin, Alexander Theus, Thomas Hofmann, Valentina Boeva, T. Konstantin Rusch, and Antonio Orvieto. Fixed-point reasoners: Stable and adaptive deep looped transformers. *arXiv preprint arXiv:2606.18206*, 2026.

Max Revay, Ruigang Wang, and Ian R. Manchester. Lipschitz bounded equilibrium networks. *arXiv preprint arXiv:2010.01732*, 2020.

Yousef Saad. *Numerical Methods for Large Eigenvalue Problems*. SIAM, 2nd, revised edition, 2011.

Lloyd N. Trefethen and Mark Embree. *Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators*. Princeton University Press, 2005.

Twan van Laarhoven. L2 regularization versus batch and weight normalization. *arXiv preprint arXiv:1706.05350*, 2017.

Ezra Winston and J. Zico Kolter. Monotone operator equilibrium networks. In *NeurIPS*, 2020.

Tong Zhang, Junhao Hu, Yun Peng, and Tao Xie. When does recurrence become an algorithm? convergence selection in weight-tied looped transformers. *arXiv preprint arXiv:2607.20594*, 2026.

A PROTOCOL AND RUN PROVENANCE

Tasks and splits. S3DIS uses 2-way 1-shot few-shot semantic segmentation under the corrected FS-PCS protocol of An et al. (2024), cross-validation fold S_1, and the validation split. A cached set of 1,500 episodes (`manual_seed=123`) is built once and reused, so evaluation of a fixed checkpoint is deterministic. ScanNet uses the corresponding fold-1 protocol. The prefix-sums tasks and checkpoint follow Bansal et al. (2022); 32-bit sequences are in distribution and 512-bit sequences are $16\times$ longer than the training length.

LoopFormer audits use fixed 1024-token blocks from a local FineWeb shard. Each block is the first 1024 tokens of a single document long enough to supply them, so no block spans a document boundary and every block carries a document id; documents shorter than 1024 tokens are skipped and counted, never padded. Four groups of 256 blocks are taken in one ordered scan, with one pair used for threshold selection and held-out testing in the allocation experiment and the other pair used for halting. Their pairwise disjointness is asserted on block hashes rather than inferred from the slicing. The blocks the preceding pilot consumed are re-derived from the earlier shard under that pilot's own concatenating loader and required not to intersect the new split. The measured overlap is 0 of 1024. These are still slices of a training shard, so absolute perplexity remains a pipeline diagnostic.

Confirmatory training group. The S3DIS runs with seeds 123, 777, and 2026 use 50 epochs, 400 episodes per epoch, `float32` training, evaluation every five epochs, learning-rate milestones at epochs 30 and 40, and the final-epoch checkpoint. Only `manual_seed` differs. The ScanNet run uses the same optimisation schedule and checkpoint rule, with the dataset, data root, and pretrained backbone changed. These four runs were selected without novel-class feedback.

Exploratory configuration group. Five older S3DIS checkpoints broaden the trajectory audit but are not controlled ablations. Plain and recall-plus-progressive complete 100 epochs. Deep-supervision-plus- clipping and clipping-only stop at epochs 49 and 72. The deep-supervision-only final checkpoint contains non-finite batch-normalisation buffers, so the earlier clean best-validation checkpoint is used; its selection involved novel-class feedback. These runs were trained with automatic mixed precision and have heterogeneous budgets. They are used only for the replicated late-dynamics observation.

Audit sampling. Each primary trajectory audit contains 100 episodes and a horizon of $T = 64$. Full per-point fields are retained for 20 episodes, Jacobian–vector-product measurements for 20, and Arnoldi estimates for 10. The projective experiment adds 100-episode, $T = 64$ geometry bundles per projective seed and a separate set of 30 shared episodes at $T = 128$ for all six depth curves. The audit runner sorts the cached episode identifiers, records the selection, and joins repeated measurements by identifier rather than position. The full-field subset is therefore the first 20 identifiers in that deterministic record. We do not treat episode index as a replicate.

Independent unit. The training run is the unit of replication. Cached episodes and the two input lengths of the prefix-sums checkpoint are repeated measurements, not independent checkpoints. For LoopFormer the checkpoint is the scope of inference and the sequence is the repeated sampling unit; three pretrained checkpoints are not independent training replications of one controlled protocol.

B DETAILED RECURRENCE AND AUDITOR

COSeg refines a correlation tensor with a stack of distinct aggregator layers. We replace that stack by one tied layer A_θ applied eight times during training. For a query scene, the persistent state is $H_t \in \mathbb{R}^{C\times(K+1)\times N}$, where N indexes query points and K is the number of foreground classes. The initial state H_0 is the unrefined correlation tensor, and the update is

$$H_{t+1} = H_t + \sigma(a)\big[A_\theta(H_t; G) - H_t\big], \tag{6}$$

where G is the fixed base guidance supplied to the original aggregator and $\sigma(a)$ is a learned scalar step size initialised at $\sigma(-2) \approx 0.12$. Confirmatory runs use the final-step episodic segmentation loss at training depth $n_L = 8$. The readout is the model's evaluation path from refined correlations

through class reduction, KPConv, and the classifier; it is captured inside the actual forward pass rather than reimplemented.

The same checkpoint is audited to $T = 64$. At every step the auditor records whole-state norm and residuals, normalised-state change, the three components in Equation (3), centred-logit change, probability Jensen–Shannon divergence, label flips, and the task metric. Per-point quantities reduce the $C \times (K + 1)$ feature dimensions. Their denominator is clamped at 10^{-8} for routine logging, but exact component attribution is restricted to non-zero point states and never uses that clamp. Whole-state identities are evaluated with `float64` reductions over the stored forward tensors.

Diagnostic precision is attached to the tensor being reduced, not to the dtype of the reported scalar. A cast of a previously stored `float32` scalar is not treated as extra information; the recurrent core must be recomputed with `float64` state storage to resolve motion below the `float32` floor.

C METRIC DEFINITIONS AND NUMERICAL RESOLUTION

The in-distribution prefix-sums column. Its terminal residual is 7.8×10^{-8}, at the `float32` floor, so Equation (1) cannot be split there: the stored norm ratio is exactly one, which bounds the norm-energy share only to $[0, 0.62]$. What the condition does establish is that the residual, the orientation term and the output changes all reach the floor together, and that r_t contracts geometrically at 0.709 per step on the way, unlike either of the other two columns.

For a whole state, we record

$$r_t = \frac{\|H_{t+1} - H_t\|_F}{\|H_t\|_F}, \qquad \rho_t = \frac{\|H_{t+1}\|_F}{\|H_t\|_F}, \qquad q_t = \left\| \frac{H_{t+1}}{\|H_{t+1}\|_F} - \frac{H_t}{\|H_t\|_F} \right\|_F .$$

The point-wise residual uses the same ratio after reducing every feature dimension except the point axis. Centered logits are $z_c = Cz$, where $C = I - K^{-1}\mathbf{1}\mathbf{1}^\top$ removes a common class-wise logit shift. We also record $\|z_{c,t+1} - z_{c,t}\|/\|z_{c,t}\|$, probability Jensen–Shannon divergence, total variation, label flips, margins, and the task metric.

Stable orientation calculation. Near $\cos\phi = 1$, subtracting $1 - \cos\phi$ loses precision: at $t = 63$ the stored $1 - \cos\phi \approx 4 \times 10^{-16}$ is two `float64` ulp of one, and carries no information. The audit therefore computes q_t directly from normalised tensors in `float64`, where the subtraction happens between tensors rather than between two nearly equal scalars. For the best shared radial fit, it also computes

$$\alpha_t = \frac{\langle H_{t+1}, H_t \rangle_F}{\|H_t\|_F^2}, \qquad p_t = \frac{\|H_{t+1} - \alpha_t H_t\|_F}{\|H_t\|_F}.$$

The scalar identity in Equation (1) is checked at every episode and step. The retained late 3D bundles predate retention of p_t, but $p_t = \rho_t \sin\phi_t$ is recovered from the retained ρ_t and q_t without cancellation, which is how Table A1 reports it. Propagating the resolution of the stored r and ρ through the identity instead gives a bound near 9×10^{-4}, about four orders weaker, because r and $\rho - 1$ agree to one ulp and their difference is noise. We do not report that route.

Current `float32` stage. One `float32` ulp at $r \approx 0.34$ is 2.98×10^{-8}. Across 100 episodes per 3D run, the terminal episode-mean residual occupies one representable value. Within each of the 20 full-field episodes, the terminal point-wise field contains at most two adjacent values on S3DIS and three on ScanNet. The supported statements are therefore cross-episode spread $\leq 2.98 \times 10^{-8}$ and point-wise CV $\leq 9.1 \times 10^{-8}$. A smaller CV computed from the stored scalars is not reported as a measurement.

The same limit applies to q. Its terminal value is 2.86–2.99×10^{-8} in all nine 3D runs across both datasets, a range too narrow, and too close to the relative spacing of the `float32` states it is computed from, to be read as nine separate measurements. We report it as an upper bound. The trajectory reaches that bound at $t \approx 55$ (Table A10); before then q is above the floor by decades and is a measurement, which is why the decay rate κ is fitted only on $t \in [16, 50]$. The recurrent core has now been rerun in `float64` on 20 matched episodes for each of three confirmatory seeds, with a `float64`-arithmetic/`float32`-storage control. All three precision modes are audited inside one host forward from one H_0 and one guidance tensor, so the episodes are matched by construction

Table A1: **Late 3D motion is shared scale drift with geometrically decaying orientation change.** "gap" is $|r_{\mathrm{raw}} - (\hat{\rho} - 1)|$ at $t = 63$; one `float32` ulp at this scale is 2.98×10^{-8}. "non-shared" is the worst episode's $\hat{\rho} q \sqrt{1 - q^2/4}/r_{\mathrm{raw}}$, the exact share of the step orthogonal to the state; it is an upper bound at the `float32` forward floor (App. C). κ is the fitted geometric decay rate of q_t over the steps where q_t is at least two decades above that floor. The trajectory-level factor μ uses a different, fit-quality-selected window and is therefore reported separately in Table A5. Columns 3–6 are terminal values over 100 episodes; provenance is in App. A.

Arm	Dataset	r_{late}	$\hat{\rho}$	gap	non-shared $\leq$	κ
Confirmatory and dataset replication						
seed 123	S3DIS S_1	0.340251	1.340251	3.0×10^{-8}	1.2×10^{-7}	0.7500
seed 777	S3DIS S_1	0.342527	1.342527	0	1.2×10^{-7}	0.7488
seed 2026	S3DIS S_1	0.342604	1.342604	3.0×10^{-8}	1.2×10^{-7}	0.7476
ScanNet fold 1	ScanNet	0.345405	1.345405	5.4×10^{-8}	1.1×10^{-7}	0.7470
Exploratory configuration robustness						
plain substrate	S3DIS S_1	0.340807	1.340807	6.0×10^{-8}	1.2×10^{-7}	0.7469
recall + progressive	S3DIS S_1	0.328339	1.328339	6.0×10^{-8}	1.2×10^{-7}	0.7531
deep sup. + clip	S3DIS S_1	0.338827	1.338827	0	1.2×10^{-7}	0.7486
clip only	S3DIS S_1	0.339426	1.339426	3.0×10^{-8}	1.2×10^{-7}	0.7472
deep sup. only	S3DIS S_1	0.345346	1.345346	3.0×10^{-8}	1.1×10^{-7}	0.7440

rather than by draw order; every bundle records the shared state digest and a pairing gate compares them before any interval is computed. This matters here because the host is not reproducible between processes. The forward mutates an exponential-moving-average prototype buffer that is not guarded by the training flag, and the neighbourhood kernels accumulate with atomics whose summation order varies per launch. The result resolves orientation change five decades below the `float32` plateau on those 60 trajectories, and eight decades below it at $T = 128$ (Table A2). It is not an end-to-end double-precision forward. The backbone, guidance construction, neighbourhood kernels, and readout remain in `float32`.

Retrieval boundary for the depth extension. The complete $T = 96$ per-episode records are retained and re-derive the primary precision result. For the separately frozen $T = 128$ extension, the three-seed summary, matched-pair checks, and reductions are retained, but full episode arrays survive only for seed 123; arrays for seeds 777 and 2026 are unavailable without rerunning. Thus the quoted $T = 128$ scalars remain inspectable but cannot be regenerated from raw arrays for those two seeds. We label the extension as secondary sensitivity evidence and anchor the storage-floor claim at $T = 96$. Its retained summary gives a plateau ratio $10^{8.44} \approx 2.8 \times 10^8$, compared with the storage-resolution ratio $2^{29} \approx 5.4 \times 10^8$.

The terminal per-point field. At the final step every fully dumped 3D episode occupies at most three adjacent `float32` residual values; for the plain S3DIS checkpoint, 19 of the 20 fully dumped episodes, holding 55,128 points in total, are bit-identical within the episode and the remaining one takes two adjacent values. This is evidence against quantization-robust late-stage heterogeneity. For a rule that halts point i when $r_t(i) < \tau$, a threshold inside that ULP-scale interval may split points, but the split is unstable to one-ULP perturbations and therefore carries no ordering.

Earlier `float16` dumps. The first audit dumps used `float16`, whose spacing near 0.34 is 2.441×10^{-4}. Every terminal value landed on one code, so the supported bounds were standard deviation $\leq 1.22 \times 10^{-4}$ and CV $\leq 3.53 \times 10^{-4}$. Promoting those stored scalars to `float64` does not restore information. They are retained only as a precision-control history; all main results use the `float32` audit.

Table A2: **On the tested trajectories the late plateau is a storage floor, and adding depth separates the two arms further.** 3 confirmatory seeds $\times$ 20 matched episodes ($n = 60$ per mode); all three precision modes are audited inside one host forward from one H_0, so the pairing is an identity rather than an assumption and every episode records a shared state digest. Values are $\log_{10} \min_t q_t$ for $t > 30$ and the late decay per step, given as the range over seeds. The paired, seed-stratified contrast at $T = 96$ is -4.917 dex, CI95 $[-4.922, -4.911]$; the arithmetic-only control moves by only -0.052 dex, which is what separates arithmetic from storage. At $T = 128$ the `float32` arm does not move at all while double precision keeps decaying, widening the contrast to -8.439 dex $[-8.452, -8.426]$: a floor is exactly what does not move when steps are added below it. † The three-seed summary, manifests and pairing checks are retained, but the retrieval archive preserves full $T = 128$ episode arrays only for seed 123; these rows are secondary sensitivity evidence and the fully retained $T = 96$ bundles anchor the claim. Scope: `float64` *recurrent core* — H_0 and the guidance come from the `float32` backbone and the readout stays `float32`, because the neighbourhood kernels have no double implementation.

		$\log_{10} \min_t q_t, t > 30$	late decay per step
$T = 96$	`float32` (as published)	-7.552–-7.550	0.9014–0.9026
$T = 96$	**`float64` recurrent core**	-12.485–-12.442	0.7449–0.7460
$T = 96$	`float64` arithmetic, `float32` storage	-7.604–-7.603	0.8991–0.9003
$T = 128^{\dagger}$	`float32` (as published)	-7.551–-7.551	0.9489–0.9496
$T = 128^{\dagger}$	**`float64` recurrent core**	-16.002–-15.969	0.7479–0.7484
$T = 128^{\dagger}$	`float64` arithmetic, `float32` storage	-7.604–-7.603	0.9476–0.9483

Table A3: **Per-point effective depth is uniform in every 3D run and heterogeneous in the controls.** "codes" is the largest number of distinct `float32` values the terminal per-point residual field takes in any one episode. Effective depth is the last step at which a point's residual is still above τ, computed from the full (T, N) field; τ is set relative to the arm's own terminal per-point residual so that the comparison does not depend on residual scale. "episodes not uniform" counts episodes whose points do not all share one effective depth at $\tau = 0.1r_T$. Twenty fully dumped episodes per arm.

Arm	codes	episodes not uniform	depths τ=$0.5r_T$	depths τ=$0.1r_T$	s.d. (steps)	points
seed 123	2	0	1	1	0.00	55128
seed 777	2	0	1	1	0.00	55128
seed 2026	2	0	1	1	0.00	55128
ScanNet fold 1	3	0	1	1	0.00	102420
plain substrate	2	0	1	1	0.00	55128
recall + progressive	2	0	1	1	0.00	55128
deep sup. + clip	2	0	1	1	0.00	55128
clip only	2	0	1	1	0.00	55128
deep sup. only	2	0	1	1	0.00	55128
contractive control	4524	19	33	20	7.27	55128
prefix-sums, in dist.	32	0	2	1	0.00	640
prefix-sums, 16×	512	20	60	58	18.88	10240

Aggregation. Unless stated otherwise, a table cell is the mean of per-episode values at $t = 63$. The norm-energy share in Table 1 is first computed within each episode as $(\rho_t - 1)^2 / r_t^2$ and then averaged. Task change is the within-checkpoint metric at $T = 64$ minus the metric at that checkpoint's training depth. These two task metrics are not compared across tasks.

Table A5 uses the same ratio-first convention for its heterogeneity-share column: $E_{\text{hetero}}/(E_{\text{hetero}} + E_{\text{direction}})$ is formed at each episode and step where both components clear the resolution gate, and those ratios are then averaged. Its three printed energy columns are averaged separately over the locations where each individual component is resolved. Consequently, the share column is a mean-of-ratios on a common mask, not a ratio-of-means recoverable from the rounded energy columns.

D Orthogonality of the three-way decomposition

For non-zero point states, Equation (2) gives

$$\sum_i (a_i - \bar{a})\|h_{t,i}\|^2 = 0, \qquad \langle \Delta_i - a_i h_{t,i}, h_{t,i} \rangle = 0.$$

Therefore

$$\langle \Delta^{\text{shared}}, \Delta^{\text{hetero}} \rangle_F = \bar{a} \sum_i (a_i - \bar{a})\|h_{t,i}\|^2 = 0,$$

$$\langle \Delta^{\text{shared}}, \Delta^{\text{direction}} \rangle_F = \bar{a} \sum_i \langle h_{t,i}, \Delta_i - a_i h_{t,i} \rangle = 0,$$

$$\langle \Delta^{\text{hetero}}, \Delta^{\text{direction}} \rangle_F = \sum_i (a_i - \bar{a}) \langle h_{t,i}, \Delta_i - a_i h_{t,i} \rangle = 0.$$

Pythagoras then yields Equation (4). Points with $\|h_{t,i}\| = 0$ cannot be assigned a radial coefficient; a full-state implementation must count them separately and place their entire update in a declared zero-state category. Adding an epsilon to Equation (2) would destroy the exact identity and is not used for mechanism attribution.

E Readout consequence of shared scaling

Let g be differentiable at a non-zero state H and invariant to positive global rescaling in a neighbourhood of one: $g(cH) = g(H)$. Differentiating this identity with respect to c at $c = 1$ gives the

familiar radial null of scale-invariant parameterisations (van Laarhoven, 2017; Arora et al., 2019):

$$\frac{d}{dc} g(cH)\Big|_{c=1} = J_g(H)[H] = 0.$$

Because $\Delta^{\text{shared}} = \bar{a}H$ for the whole tensor, $J_g(H)[\Delta^{\text{shared}}] = 0$. Taylor expansion therefore gives

$$g(H + \Delta) - g(H) = J_g(H)[\Delta^{\text{hetero}} + \Delta^{\text{direction}}] + O(\|\Delta\|_F^2)$$

whenever the three-way decomposition is defined. If the readout is exactly invariant for the finite positive scale $1 + \bar{a}$, then the shared-only finite perturbation is also an exact null. Neither conclusion applies automatically to the heterogeneous or directional components.

This result is deliberately conditional on the readout. A homogeneous linear logit head without normalisation can map a positive state scaling to a positive logit scaling. The argmax is unchanged, while softmax probabilities and calibration may change. Conversely, a normalisation immediately before the head can remove the shared mode. The position-wise LoopFormer intervention in App. H distinguishes those cases; it is the empirical test of the assumption, not a consequence inferred from geometry alone.

F Additional controls

Synthetic operator. A two-layer MLP performing kNN label propagation on synthetic 3D tasks reproduces the combination of state-norm growth, near-collinear motion, probability saturation, and depth degradation. Its terminal per-point residual CV remains 0.0198–0.0959, however, so the point-wise homogenisation seen in the COSeg-based refiner is not a generic consequence of weight tying. The synthetic operator is an instrument sanity check.

Matched contractive control. One further S3DIS checkpoint recurs on a prediction state rather than on the correlation tensor. It was trained by the same command line as the seed-123 confirmatory run with only the configuration file changed: same 50 epochs, same learning-rate schedule, same 400 episodes per epoch, same `float32` precision, same fixed-epoch checkpoint rule, both completing without non-finite values. It is then audited by the same probe, horizon, and 100 cached episodes.

It contracts. The norm ratio is 1.000003, r falls from 9.5×10^{-2} to 7.6×10^{-6} at a fitted rate of 0.867, the terminal per-point field takes $4{,}524$ distinct `float32` values in an episode instead of two, its points spread over 20–33 effective depths (Table A3), and its task curve peaks at $t = 14$ — after the training depth — and then loses 0.0011 rather than 0.0265. Final novel-class validation mIoU differs by 0.006 between the two.

This serves two purposes. It is the instrument control for Section 4.2: within one task and one codebase, the same measurement reports uniformity for the nine audited runs and wide heterogeneity here, so the uniformity is a property of those runs. It is also a within-protocol contrast, unlike the damping control below, but changing the persistent-state definition changes several downstream dynamics. It remains one seed pair, so it does not estimate a run-level effect of that design choice.

The contractive control as a positive control for the decomposition. Tiny reported non-shared components could reflect either genuine dynamics or limited sensitivity of the probe. The same probe, run on the contractive checkpoint above, returns $E_{\text{hetero}} = 0.211$ and $E_{\text{direction}} = 0.666$ — four to five orders above the 3D arms' 5×10^{-6} and 1.6×10^{-3}. The instrument resolves order-one heterogeneity when it is there. Its $E_{\text{shared}} = 0.123$ is specific to this state definition. This arm's state is a per-point probability vector, whose radial direction is different from the scale mode of the correlation tensor. Accordingly, the row only reports the two orthogonal components. This arm also carries one of the two fully valid deletion rollouts in the paper. All three components can be removed while the terminal norm stays within 16% of the sham's, making Tier D interpretable here. The prefix-sums checkpoint supplies the other such system.

Projective state (single-flag patch). The Section 4.5 arm inserts three lines into the tied update, gated by one configuration flag: after the block computes its output, the output is rescaled to the input's Frobenius norm. The rescale is inside the step function the auditor drives, not around the

loop that calls it, so the audited trajectory and the trained one are the same dynamical system. With the flag off, the model executes the unchanged update path. Everything else — the command line, the 50-epoch schedule, the 400 episodes per epoch, the `float32` precision, the fixed-epoch checkpoint rule, and the episode cache used for evaluation — is shared with the confirmatory runs, and the three projective seeds and three control seeds are evaluated on one cache built for this round.

The resolved configuration stored in each $T = 128$ evaluation JSON is a snapshot of the mutable configuration file at audit time, not the training manifest. It therefore contains later defaults such as `epochs=100` and `sel_protocol=novel_oracle`; the checkpoint manifests and training logs record the actual 50-epoch, `fixed_epoch` protocol. Selection provenance in this paper comes from those immutable records, never from the evaluation-time snapshot.

The patch is a reimplementation. The original was written on a machine that was decommissioned before the edit was retrieved, so the code behind Section 4.5 was written again from the recorded configuration and validated before it was used: the surviving checkpoint from the original run was replayed through it on 100 episodes, and the recorded terminal geometry came back — $\rho = 0.9999999979$ against a recorded 1.0000000042, $E_{\text{direction}}$ within 0.2%, and μ on the *same* 16-step window the original selected, which is chosen by fit quality and never inspects μ. One degree of freedom cannot be recovered by forward replay. It is unknown whether the norm ratio was inside the autograd graph during the original training. Seed 123 was therefore retrained under the reimplementation, and its per-epoch training curve was compared against the original's. The mean absolute difference falls from 0.039 over the first half of training to 0.005 over the second. This is consistent with, but does not prove, a matching gradient path; no claim depends on the unrecovered original run. All three runs reported in Section 4.5 are new, and none of the recorded original run's numbers enter any claim.

One-seed damping control. Within the 3D architecture, changing the configured damping constant from 0.9 to 0.2 moves the measured norm ratio from 2.23 to 1.008 and raises terminal pointwise residual CV from below the resolution bound to 0.1225. The residual at $t = 31$ is 0.0126 and is still decreasing. Both settings were trained under automatic mixed precision and stopped early — at epochs 100 and 76 — so they are not on the `float32` confirmatory protocol, each has one separately trained checkpoint, and the broader configuration also changes evidence reinjection and spectral normalisation relative to the plain refiner. This control shows that the measured regime is not architecturally inevitable; it does not isolate its cause or establish a remedy.

G Supplementary tables from the main audit

The main text reports the cross-seed ranges and decision-relevant contrasts. The full component, intervention, and depth-landmark tables are retained here so every printed claim remains traceable without consuming the initial-submission page budget.

H Where the shared-scale null lives

A zero flip rate under a shared-scale injection admits two interpretations. Either the state carries no functional scale information, or the readout removes scale before anything downstream can see it. Injecting at three positions on the same sequences separates them, with λ matched to the norm of whichever tensor is perturbed, so the absolute magnitudes differ between positions by construction (Table A7).

Before the RMS normalisation the shared mode moves the centred logits by 4.0–4.5×10^{-7}, which is the `float32` arithmetic floor (3.6×10^{-7}), and shifts NLL by less than 10^{-8} nats. Applied to the normalised activation one module later, the identical perturbation moves the centred logits by exactly λ and shifts NLL by $+1.17$ to $+1.71$ nats — a ratio of about 1.1×10^{6}. The observed centred-logit change of 0.5 matches the analytic prediction for shared rescaling through a linear head, supporting the intervention implementation. So it is a property of where the normalisation sits, not evidence that the state carries no scale information.

Two boundaries. First, the flip rate is 0 even with the normalisation ablated, where ΔNLL exceeds 9 nats: a positive rescaling cannot move an argmax, so a null flip rate is never sufficient evidence

Table A4: **The probe separates state geometry from readout calibration in a modern looped Transformer.** Three released LoopFormer checkpoints, matched at 24 block-equivalent evaluations, audited with the identical instrument on 256 held-out sequences each and reported per checkpoint rather than pooled. All three sit in the regime opposite to the 3D refiner — orientation rather than scale, and output change remains at the audited horizon. Half a state norm injected along the global shared-scale mode flips no token, consistent with the checkpoint's pre-readout RMS normalisation; the same energy along a directional mode flips most of them. Appendix H localises that null to the normalisation rather than to the state. "flip" columns are bootstrap mean flip rate at amplitude 0.5; a zero-amplitude injection returns exactly zero on every sequence. Attaching the auditor leaves the logits bit-identical on 3 batches per checkpoint. Perplexity is a diagnostic: the release publishes no per-checkpoint validation number, so the reproduction status is `reproduction-unverified`.

					last flip	flip at $\lambda = 0.5$	
	ppl	r_T	q_T	E_{shared}	/ iters	shared	direction
3 blk × 8	22.83	0.996	1.073	0.513	8/8	0.0000	0.751
2 blk × 12	24.16	1.001	1.058	0.424	12/12	0.0000	0.744
1 blk × 24	27.47	0.724	0.716	0.126	24/24	0.0000	0.634

Table A5: **The non-shared part of the step is directional, not per-point scale.** Energies are episode-then-run means over $t \in [8, 30]$, the window where both small components clear three rounding floors; steps outside it are excluded rather than averaged in. The three energy columns average each component over the locations where that component is resolved. The heterogeneity-share column instead forms $E_{\text{hetero}}/(E_{\text{hetero}} + E_{\text{direction}})$ first at each episode and step where both components are resolved, then averages those ratios; it is therefore not the ratio of the two printed mean-energy columns. $\mu = \kappa\bar{\rho}$ is measured on the 16-step window whose $\log q$ fit is straightest, a criterion that never inspects μ (App. M). The contractive control is audited by the identical probe and shows what the same code reports when the small components are large; its E_{shared} is not comparable to the 3D rows, because its state lives in the probability simplex.

Arm	E_{shared}	E_{hetero}	$E_{\text{direction}}$	$\frac{E_{\text{hetero}}}{E_{\text{hetero}}+E_{\text{direction}}}$	μ	window
Confirmatory seeds (differ only in `manual_seed`)						
seed 123	0.998401	5.0×10^{-6}	1.6×10^{-3}	0.0021	1.00224	34–49
seed 777	0.998528	9.8×10^{-6}	1.5×10^{-3}	0.0047	1.00262	34–49
seed 2026	0.998466	7.9×10^{-6}	1.5×10^{-3}	0.0034	1.00162	33–48
Exploratory configurations						
plain	0.994760	4.1×10^{-5}	5.2×10^{-3}	0.0056	1.00065	33–48
recall + prog.	0.997429	2.2×10^{-6}	2.6×10^{-3}	0.0013	0.99999	31–46
deep sup. + clip	0.997090	1.3×10^{-5}	2.9×10^{-3}	0.0024	1.00246	30–45
clip only	0.995947	1.2×10^{-5}	4.0×10^{-3}	0.0024	1.00005	32–47
deep sup. only	0.997252	6.5×10^{-6}	2.7×10^{-3}	0.0023	1.00044	31–46
Contractive control, same task and codebase						
contractive control	0.122585	2.1×10^{-1}	6.7×10^{-1}	0.2418	0.87659	47–62

of invisibility, and every position here is reported on probabilities and NLL as well. Second, this localises the null. It does not show that the natural shared drift mediates anything, and the ablated readout is an ablation of these weights rather than a second architecture, so it licenses no cross-family claim.

I THE EQUAL-COST RETIMING CONTROL

Section 4.6 compares three schedules that all cost k iterations × blocks-per-iteration. *native-k* passes `steps=` $[1/k]^k$ and integrates internal time to 1. *truncated-k* passes the full fine grid $[1/N]^N$ and stops after k, integrating time only to k/N; this is what a residual threshold bolted onto the fine grid actually produces. *retimed-k* passes $[1/N]^{k-1}$ followed by a single step of size $1 - (k-1)/N$:

Table A6: **At matched cost the model's own schedule dominates every tested dynamic rule, and an equal-cost control explains why.** Change in next-token NLL (nats; lower is better) against the full budget, on a 256-sequence slice touched once after the thresholds were frozen. "Own budget" uses the native k-step schedule; "retimed" uses the first $k-1$ fine-grid steps and a final step that completes the internal clock, at the same k-step cost as truncation; "truncated" stops after k steps on the fine grid and leaves the clock unfinished. "best rule" is the best of four frozen thresholds, applied to q, the raw residual, Jensen–Shannon divergence, and centred-logit change, among those whose realised cost mismatch stayed within the frozen 2 percentage-point tolerance (largest realised mismatch 1.66 points). The final column is the paired bootstrap interval for that rule minus the native budget. Retiming recovers 85%–97% of the truncation gap (median over budgets, per checkpoint), so the gap is dominated by the unfinished time integral rather than by schedule coarseness. The retiming comparison is a mechanism diagnostic; it explains the gap the frozen test measures and is not itself a halting claim.

	less compute	own budget	retimed	truncated	best rule	paired CI (rule − own)
3 blk × 8	25%	+0.0264	+0.0916	+3.2604	+1.6720	[+1.399, +1.892]
3 blk × 8	50%	+0.0947	+0.2207	+3.7828	+3.4906	[+3.231, +3.552]
2 blk × 12	25%	+0.0275	+0.1425	+2.7907	+2.7345	[+2.634, +2.773]
2 blk × 12	50%	+0.0835	+0.3134	+3.3499	+2.9391	[+2.637, +3.058]
1 blk × 24	25%	+0.0278	+0.4403	+2.3123	+1.9451	[+1.792, +2.040]
1 blk × 24	50%	+0.1252	+0.6918	+2.9701	+3.4138	[+3.128, +3.456]

Table A7: **The shared-scale null is a property of where the normalisation sits, not of the state.** The same equal-energy injection at $\lambda = 0.5$, read at three positions on 256 held-out sequences per checkpoint; λ is matched to the norm of whichever tensor is perturbed, so the absolute magnitudes differ between positions by construction and are recorded separately. Before the RMS normalisation the shared mode moves the centred logits by about 3.96×10^{-7}, which is the `float32` arithmetic floor (3.58×10^{-7}); one module later the identical perturbation moves them by exactly λ and shifts NLL by more than a nat — a ratio of about 1.1×10^{6}. The exact 0.5 is the analytic value: a shared-mode injection is a rescaling, and a linear readout scales centred logits by $1 + \lambda$. The flip rate is 0 even with the normalisation removed, where ΔNLL exceeds 9 nats, which is why a null flip rate is never read here as invisibility. The ablated-normalisation row is an ablation of the released readout on the same weights, not a second architecture, and licenses no cross-family claim.

		shared			direction	
	injection position	flip	centred-logit	ΔNLL	flip	ΔNLL
3 blk × 8	pre-norm state, released readout	0.0000	4.55×10^{-7}	-1.86×10^{-9}	0.751	+0.368
	post-norm activation, released readout	0.0000	5.00×10^{-1}	+1.656	0.470	−1.191
	pre-norm state, normalisation ablated	0.0000	5.00×10^{-1}	+12.031	0.523	+2.377
2 blk × 12	pre-norm state, released readout	0.0000	4.39×10^{-7}	-1.30×10^{-8}	0.744	+0.325
	post-norm activation, released readout	0.0000	5.00×10^{-1}	+1.712	0.446	−0.909
	pre-norm state, normalisation ablated	0.0000	5.00×10^{-1}	+10.856	0.587	+0.994
1 blk × 24	pre-norm state, released readout	0.0000	3.96×10^{-7}	7.45×10^{-9}	0.634	+0.104
	post-norm activation, released readout	0.0000	5.00×10^{-1}	+1.173	0.464	−0.263
	pre-norm state, normalisation ablated	0.0000	5.00×10^{-1}	+9.449	0.519	+1.736

identical to truncated-k for its first $k-1$ steps, identical to native-k in finishing the clock, and identical to both in cost. The three therefore isolate the time integral from the shape of the schedule.

The fraction of the truncated-versus-native gap that retiming recovers is 96.2–100% on the 3-block checkpoint, 92.5–100% on the 2-block and 80.4–100% on the 1-block, over budgets from a 25% to a 96% saving; the medians are 97.5%, 94.6% and 85.5%. The 100% endpoints occur at $k = 1$, where retimed-1 and native-1 are the same schedule by construction — a built-in check that the two objects are what they claim to be, and the reason the main text quotes medians rather than the raw

Table A8: **A per-token settling depth is not identified on this family, which is a different statement from the depths being equal.** Threshold on q selected on a disjoint validation slice and frozen; the held-out slice was entered once. Right censoring stays below the pre-specified 0.80 bar, but the depth spread is zero and sits exactly on a token-identity permutation null, so no token-specific structure is present at these thresholds. The re-run of the withdrawn self-referential rule censors 1.0000 of tokens, which is why that earlier number was uninterpretable rather than merely imprecise. The fifth column applies the convention the other two families already use (τ a fraction of the *episode-mean* terminal residual, one threshold shared by all units), restoring a like-for-like cell: it gives 1 on the nine 3D arms, 20 on the contractive OSRR control and 58 on extrapolating prefix-sums. The last column is why: the per-token residual ends far above the smallest value it reached, whereas the 3D per-point residual ends exactly at its minimum (ratio 1.0000). The trajectory never settles, so the quantity a settling rule estimates is undefined here.

	frozen τ_q	censored fraction	IQR (depths)	null 97.5% IQR	published-conv. distinct depths	last/min ratio
3 blk $\times$ 8	1.103	0.441	0.000	0.000	1–1	8.65
2 blk $\times$ 12	1.075	0.449	0.000	0.000	1–1	14.54
1 blk $\times$ 24	0.734	0.439	0.000	0.000	1–1	15.66

range. This curve was computed on the validation slice: the held-out slice is entered once, by the frozen test, and a diagnostic added afterwards must not reuse it.

J Component-deletion rollouts

Deleting a component for an entire rollout is the stronger test, and on the 3D arms it is not available. The shared component is 99.8% of the step, so removing it for fifty steps leaves the terminal state at 10^{-8} of the sham's norm. The resulting task drop of 0.18–0.34 is therefore not a measurement of what the component does. It is a measurement of a readout evaluated eight orders of magnitude outside its operating range, and we report it as one (Table A9). Holding the global norm fixed instead does not rescue the test. It costs *more* (0.33–0.48) because the operator is not scale-equivariant. Dividing out the scale gives a different dynamical system rather than the same trajectory rescaled (App. N). Only the heterogeneity deletion stays on the manifold on these arms, at 0.985–0.990 of the sham's norm, and it costs 0.0004–0.019.

Two systems where all three deletions do stay on the manifold provide complementary checks. On the contractive control, deleting the shared component changes the terminal task by $+0.0000$ and deleting the directional component costs 0.164. On the pretrained prefix-sums checkpoint at $16\times$ its training length, norm clamping moves accuracy by $+0.002$ while deleting direction costs 0.093. At its training length, the same model already achieves 100% accuracy, so the absence of an intervention effect is ceiling-limited rather than evidence of insensitivity. In both on-manifold systems, directional deletion has the largest adverse task effect. These controls support the distinction between component energy and functional influence, but they do not make the off-manifold 3D deletions valid.

The complete intervention matrix, including off-manifold diagnostics, is deferred to Table A9 in App. G.

K Full projective-training analysis

We retrained the confirmatory configuration with one change: after each tied update the state is rescaled to its pre-update norm. The command line and schedule are otherwise unchanged, with the intervention gated by one configuration flag. Three seeds, against the three control seeds, are all evaluated on one episode cache (Table A11).

The constraint does what the geometry asks for in every seed. Holding the norm forces $\rho_t = 1$ by construction, while the orthogonal mode contracts in all three runs: $\mu = 0.923$–0.991 against the controls' 1.00162–1.00262. The step energy also inverts from $E_{\text{shared}} > 0.998$ to $E_{\text{direction}} =$

Table A9: **Deleting a component for the whole rollout: task change against the sham.** Closed-loop counterfactuals from $t = 0$, decomposed afresh at every step, 20 episodes per arm, paired within episode. † marks an arm whose terminal state norm ends more than a decade away from the sham's; the readout is then outside the range it was calibrated on and the number measures that, not the component. On the 3D arms the shared component *is* the step, so deleting it cannot be tested this way – which is why the two systems in the lower block, where all four interventions stay on the manifold, carry the comparison. ‡ marks an arm at 100% accuracy, where no intervention can move the metric in either direction; those zeros are a ceiling, not a null result. Norm clamp holds the global state norm at its value when the rollout starts; it is not the same as deleting the shared component, since it also absorbs the norm change the other two produce.

Arm	delete shared	delete hetero	delete direction	norm clamp
Confirmatory 3D seeds				
seed 123	-0.1834†	-0.0004	-0.0326†	-0.3255†
seed 777	-0.2394†	-0.0058	-0.0137†	-0.4821†
seed 2026	-0.3436†	-0.0185	-0.0079†	-0.4250†
Systems where every deletion stays on the manifold				
contractive control	+0.0000	-0.0008	-0.1636	-0.0000
prefix-sums, 16× length	-0.0166	-0.0479	-0.0928	+0.0020
prefix-sums, trained length	+0.0000‡	+0.0000‡	+0.0000‡	+0.0000‡

Table A10: **The task curve settles and the labels freeze before the step becomes pure scale drift.** Depths are read from episode-mean curves by fixed rules: task peak is the first maximum; task settle is the first depth after which the metric stays within 0.002 of its value at $T = 64$; label freeze is the first depth after which no predicted label changes again; the two regime columns are the first depths after which the norm-energy share $(\rho - 1)^2/r^2$ stays above 0.99 and $1 - 10^{-6}$; q floor is the first depth at which the normalised-state change reaches twice its terminal value, and is left blank where q has not flattened by $T = 64$. "Δ" is the task metric at $T = 64$ minus its value at the peak, in that column's own units. "–" means the criterion is never met within $T = 64$. Where Δ is zero the peak location carries no information: the whole curve after "settle" lies inside the margin.

Arm	train	task peak	Δ	task settle	label freeze	$E_{\mathrm{norm}}{>}0.99$	$E_{\mathrm{norm}}{>}1{-}10^{-6}$	q floor
seed 123	8	5	-0.0265	19	39	9	26	56
seed 777	8	5	-0.0377	16	43	9	25	55
seed 2026	8	1	-0.0417	13	41	9	25	55
ScanNet fold 1	8	21	0.0000	6	40	9	26	55
contractive control	8	14	-0.0011	10	62	–	–	58
prefix-sums, in dist.	30	14	0.0000	13	14	–	–	52
prefix-sums, 16×	30	64	0.0000	64	–	–	–	–

0.986–0.994. These are realised trajectory dynamics, not a global contraction certificate for the operator.

At the training depth, the three runs do not show a consistent signed task difference. At $n_L = 8$, full-resolution validation mIoU on the shared episodes gives paired projective-minus-control differences of -0.009, $+0.029$, and -0.020. With three training runs per condition, these values neither establish equivalence nor support a systematic penalty or benefit. A single-run version of this experiment reported a 0.021 mIoU penalty. That was one draw, and it does not survive three seeds a side.

The low-resolution depth-curve landmarks show the more robust separation. Using the pre-specified IoU margin of 0.002, the control seeds settle at $t = 13$, 15 and 18; the projective seeds settle at $t = 61$, 107, and 121. From $t = 32$ to $t = 128$, every control seed is unchanged to four decimals, with largest movement 1.2×10^{-5}. Every projective seed peaks at $t = 7$ and then falls to 0.126, 0.217, and 0.033. Peak-to-terminal loss is 0.208–0.420 for the projective arms against 0.058–0.098 for the controls. Read from $t = 0$ rather than from the training depth, the two conditions differ in both shape and level: every control seed peaks at $t \in \{1, 2, 4\}$ and is already declining by the depth

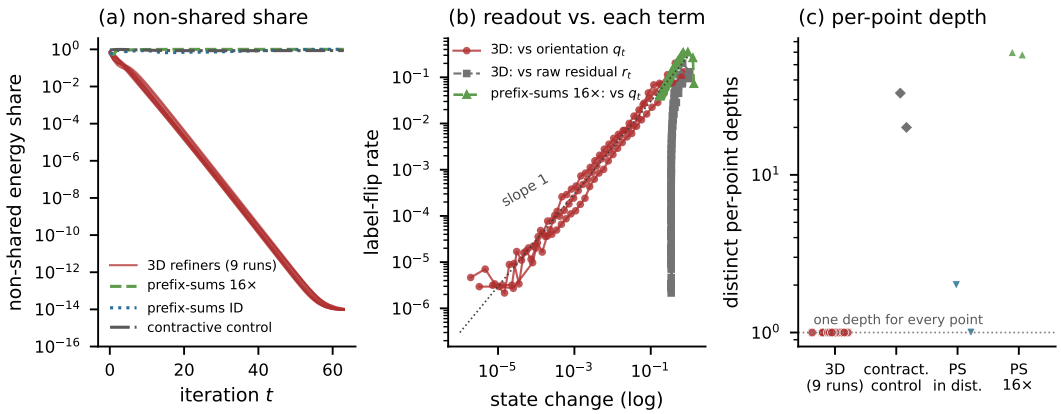

Figure A1: **Depth-wise and unit-wise checks.** (a) The non-shared energy share falls to the `float32` floor in all nine 3D runs, but remains large in prefix-sums and the contractive control. (b) Label changes track q_t over several decades and not the nearly constant raw residual. (c) The tested per-point effective-depth rule returns one depth for the audited 3D runs, but many depths for the positive controls.

it was trained at, while every projective seed climbs monotonically to $t = 7$. Each projective run peaks near its training depth, then loses 0.208–0.420 IoU by $t = 127$.

The second is that contraction does not buy a halting signal. At $\tau = 0.1r_T$, all three projective seeds place every point in each of 20 full-field episodes at one effective depth, as the nine audited arms do. At $\tau = 0.5r_T$, two seeds do the same and the third reaches five distinct depths; the matched contractive control spreads over 20–33. Imposing norm preservation and obtaining $\mu < 1$ is therefore not sufficient to make the per-point residual informative.

The output pattern replicates. Every control seed reaches exactly zero Jensen–Shannon divergence by the horizon and stops flipping labels between $t = 29$ and $t = 99$. No projective seed saturates: all are still flipping labels at $t = 127$, with terminal divergences of 2×10^{-6} to 3×10^{-5}. This ordering is consistent with saturation stabilising the controls, but does not establish saturation as a mediator of the intervention.

Across three runs per condition, the projective arms show greater observed spread in μ and late residual than the controls. In the projective runs, μ spans 0.068 versus 0.0010 across the three matched controls; r_{late} spans a factor of 7.4 versus 1.05 across the nine audited control arms. With three runs the variance contrast is only a descriptive observation. Its source is not identified.

The local injections show that the shared direction strongly changes logits but has little influence on probabilities or labels at the audited states. Projective training shows that removing norm growth can make the realised state dynamics contractive without making predictions stable under additional depth. Because the constraint also changes the training trajectory, learned weights, and subsequent directional dynamics, it does not prove that absence of the scale mode itself causes the depth degradation. It rules out the naive prescription "make the state converge."

L Expanded limitations

The evidence covers three model families. Only the 3D family has multiple independently trained checkpoints, and only three S3DIS seeds share the same confirmatory protocol. The ScanNet result is one weak checkpoint and supports a dynamics claim only. The prefix-sums and LoopFormer comparisons are checkpoint-conditional. LoopFormer trajectory and intervention results now use 256 held-out 1024-token sequences per checkpoint with sequence-level bootstrap intervals, but three pretrained checkpoints remain three instances rather than replications of one controlled protocol, so nothing is pooled across them and no population claim is made about looped Transformers in general.

Table A11: **Projective training enforces norm preservation but does not stabilise additional depth.** The projective update rescales the state to its pre-update norm; the training configuration differs from the control by that flag, but the learned dynamics can change downstream. Every seed has $\hat{\rho} = 1$ by construction and a fitted trajectory-level factor $\mu < 1$, yet every seed loses the control's depth stability. The full-resolution validation mIoU differences at $n_L = 8$ cross zero across the three run pairs and do not establish equivalence. In 20 full-field episodes per seed, $\tau = 0.1r_T$ always gives one per-point effective depth; at $\tau = 0.5r_T$, two seeds have one and one seed has five, against 20–33 for the contractive control of Section 4.2. All values use the same 30 episodes and cache. The final block is low-resolution IoU and must not be numerically compared with the full-resolution validation rows above it.

	seed		
	123	777	2026
control			
μ	1.00224	1.00262	1.00162
r_{late}	0.3413	0.3418	0.3428
full-res. val. mIoU at $n_L = 8$	0.3249	0.2670	0.3122
last step with a label flip	32	99	29
projective			
μ	0.99124	0.94053	0.92310
r_{late}	0.05392	0.00911	0.00728
$E_{\text{direction}}$ (window)	0.9919	0.9942	0.9856
full-res. val. mIoU at $n_L = 8$	0.3160	0.2963	0.2917
last step with a label flip	127	127	127
distinct per-point depths, $\tau = 0.5r_T$	1	1	5
distinct per-point depths, $\tau = 0.1r_T$	1	1	1
low-resolution IoU versus depth, $T = 128$			
control $t = 8$	0.4240	0.3719	0.3812
control $t = 16$	0.3982	0.3425	0.3668
control $t = 32$	0.3953	0.3413	0.3665
control $t = 64$	0.3953	0.3413	0.3665
control $t = 96$	0.3953	0.3413	0.3665
control $t = 127$	0.3953	0.3413	0.3665
projective $t = 8$	0.4343	0.4185	0.4459
projective $t = 16$	0.3112	0.2929	0.2997
projective $t = 32$	0.2277	0.1903	0.1712
projective $t = 64$	0.1642	0.2167	0.0473
projective $t = 96$	0.1584	0.2169	0.0365
projective $t = 127$	0.1256	0.2169	0.0331

LoopFormer blocks are now document-bounded: each is the first 1024 tokens of one sufficiently long document, so no block spans a boundary. The four 256-block groups are pairwise disjoint on block hashes and share no block with the earlier pilot, which is re-derived under its own loader and checked rather than assumed. They remain slices of a training shard, not official validation and test splits, so a held-out result here tests a frozen local rule without establishing dataset-level generalisation. Provenance now pins the immutable commit SHA of each checkpoint together with SHA-256 digests of every weight, config and tokenizer file.

The withdrawn LoopFormer allocation statistic is not reported as evidence. It defined each token's threshold as a fraction of that same token's terminal residual; at the terminal step the residual necessarily exceeds such a threshold, so every observation is right-censored at the horizon by construction, and re-running it censors 1.0000 of tokens. Its replacement selects a global threshold on a disjoint slice, freezes it, and reports censoring rates, the full threshold sweep and a token-identity permutation null. That replacement returns no identifiable per-token depth signal. Because the per-token residual ends well above its minimum, the data leave per-token settling depth unidentified for this model family.

The separation of 3D per-point scale heterogeneity from within-point directional motion holds only where both components clear the rounding floor, $t \in [8, 30]$. Later than that the two terms together

are at the `float32` forward floor and their ratio is not a measurement; the reported energies are window statistics and are not extrapolated past the window. The recurrent-core `float64` measurement now covers 20 matched episodes for each of three confirmatory seeds at $T = 96$ (60 per precision mode), with a `float64`-arithmetic/`float32`-storage control and a separately frozen $T = 128$ extension. All three modes are audited inside one host forward from one H_0, so episodes are paired by construction and the interval is an episode-paired, seed-stratified one rather than a range. It is still not an end-to-end double-precision forward. The seed remains the unit of replication. Accordingly, three seeds bound how far the result generalises across training runs, regardless of the within-seed precision.

The $T = 96$ episode arrays are fully retained. For $T = 128$, full arrays survive for seed 123, while the complete three-seed summaries, pairing checks and reductions survive for all seeds. Consequently the reported $T = 128$ numbers can be inspected but not regenerated from raw arrays for seeds 777 and 2026 without rerunning. They are secondary sensitivity evidence, while the primary storage-floor conclusion rests on the main analysis.

On the 3D arms only the heterogeneity deletion stays on the manifold. Deleting the shared component, or clamping the norm, leaves the state orders of magnitude from the sham, so those rows measure a readout outside its operating range rather than component effects. The single-step equal-energy injections avoid closed-loop compounding. They remain finite perturbations, reaching up to half the state norm, and may fall off the data manifold. They identify output sensitivity to a specified perturbation. In LoopFormer, the shared intervention is a positive global rescaling immediately before an RMS-normalised readout and is therefore expected to vanish; injecting on either side of that normalisation localises the null there rather than in the state, which is a statement about this readout structure and not independent cross-family mechanism evidence.

The growth factor μ assumes a locally linear step and describes realised growth along one trajectory, not a certified eigenvalue. Its window is selected by a stated rule whose sensitivity is tabulated in App. M. Valid 3D and prefix-sums effective depth is measured at two arm-level thresholds; other thresholds and a full first-hitting sweep are untested.

The projective comparison has three independent training runs per condition. Its ranges do not estimate a population frequency or license equivalence, and the signed trained-depth differences cross zero. The mechanism checkpoints are retrained versions that pass a pre-specified geometric gate; task landmarks from the two non-reproducible episode caches are never compared (App. Q). Projective rescaling changes learned weights, later directional dynamics, and norm growth, so it evaluates a design constraint rather than mediation by the original scale component.

Most trajectory and cross-model audits are limited to $T \leq 64$; the recurrent-core confirmatory audit reaches $T = 96$, and its secondary sensitivity extension and the projective comparison reach $T = 128$. The data show no raw-state fixed point in the expansive controls within those horizons, not that none exists. Exploratory Arnoldi estimates do not identify spectral multiplicity or a largest singular-value gain, and the cross-run predictor study is confounded by the single external-family checkpoint (App. O–P).

M CHOOSING THE WINDOW FOR μ

$\mu = \kappa\hat{\rho}$ is fitted, and a fitted number needs a stated rule for where it is fitted. It matters here more than usual, because μ drifts monotonically with window placement. On seed 123, sliding a 16-step window along the trajectory gives

window	μ	R^2 of the $\log q$ fit
$[8, 23]$	1.01881	0.9999839
$[16, 31]$	1.01015	0.9999675
$[24, 39]$	1.00482	0.9999905
$[34, 49]$	1.00224	**0.9999994**
$[44, 59]$	1.04694	0.9855249

A rule that picked the window after seeing μ could deliver any answer in $[1.002, 1.047]$. We therefore select by fit quality alone: the reported window is the one whose $\log q$ regression is straightest, and

the selection code never reads μ. On the three confirmatory seeds this lands on $[34, 49]$, $[34, 49]$, and $[33, 48]$, all with $R^2 \geq 0.9999993$, giving $\mu \in [1.00162, 1.00262]$.

The rule is not doing hidden work. It selects the same region that an a-priori rule would, past the transient and before the trajectory enters the rounding floor. The last row above shows what happens when it does not. The $[44, 59]$ window overlaps the numerical floor, lowering R^2 to 0.9855 and shifting μ to 1.04694; this window is therefore excluded from the reported estimate. The early windows are honest fits to a genuinely faster early decay, which is why they are straight too; they describe the transient rather than the asymptotics.

What survives every admissible choice is the finding the paper uses: while q traverses more than five decades, the inferred trajectory-level factor approaches 1 from above. No admissible window makes that realised factor contractive; this is not a global statement about every direction of the operator.

N THE OPERATOR IS NOT SCALE-EQUIVARIANT

Because the shared component is 99.8% of the late step, the obvious response is to divide the scale out — rescale the state to a fixed norm at each step and read the remaining dynamics. Scale equivariance, $F(cH) = cF(H)$, would make this rescaling compatible with the original dynamics; the implemented operator lacks this property.

The implemented operator contains additive guidance, normalisation layers, and a softmax inside the aggregator; its functional form is therefore not positively homogeneous in H. Normalising the state produces a *different* dynamical system rather than a rescaled version of the same trajectory. The norm-clamp rollout confirms that the altered system does not recover the sham task curve (Table A9), and the projective model of Section 4.5 is consequently trained with the rescaling inside its update rather than obtained by post-hoc normalisation. A previous diagnostic run quantified the non-equivariance but did not save its bundle; those unreproducible numbers are omitted.

O EXPLORATORY FINITE-KRYLOV DIAGNOSTICS

We apply matrix-free Arnoldi to the local step Jacobian using autograd Jacobian–vector products. For an Arnoldi factorisation $AQ_m = Q_m H_m + h_{m+1,m} q_{m+1} e_m^\top$ and a unit eigenvector y_i of H_m, the Ritz vector $u_i = Q_m y_i$ has absolute residual

$$\varepsilon_i = \|Au_i - \theta_i u_i\| = |h_{m+1,m}|\,|e_m^\top y_i|.$$

The reported relative residual is $\varepsilon_i / |\theta_i|$. Normality is not established, so this residual is interpreted as pseudospectral evidence rather than an eigenvalue-error bound. It is not a singular-value estimate and is not called a local gain (Saad, 2011; Trefethen & Embree, 2005; Golub & Van Loan, 2013).

Table A12: **Leading Ritz estimates at $m = 24$.** Magnitudes are computed at step 48 with autograd Jacobian–vector products and averaged over ten episodes. These are finite-Krylov estimates at one subspace dimension, not identified eigenvalue multiplicities or singular-value gains. Table A13 reports the dimension sweep and the relative Ritz-pair residual.

Arm	$\lvert\theta_1\rvert$	$\lvert\theta_2\rvert$	$\lvert\theta_3\rvert$	$\lvert\theta_4\rvert$	$\hat{\rho}$	$\lvert\lvert\theta_1\rvert - \hat{\rho}\rvert$
seed 123	1.340251	1.340251	1.340251	1.016475	1.340251	1.7×10^{-8}
seed 777	1.342527	1.342527	1.342527	0.987518	1.342527	1.7×10^{-8}
seed 2026	1.342604	1.342604	1.342603	0.997361	1.342604	1.4×10^{-8}
ScanNet fold 1	1.345405	1.345405	1.345359	1.071798	1.345405	5.2×10^{-8}
plain substrate	1.340807	1.340807	1.340807	1.340807	1.340807	2.4×10^{-8}
recall + progressive	1.328339	1.328339	1.328339	1.328339	1.328339	7.9×10^{-9}
deep sup. + clip	1.338827	1.338827	1.338827	1.338827	1.338827	8.7×10^{-8}
clip only	1.339426	1.339426	1.339426	1.339426	1.339426	1.7×10^{-8}
deep sup. only	1.345346	1.345346	1.345346	1.345346	1.345346	1.4×10^{-8}

Table A13: **Krylov-dimension sensitivity of the leading Ritz estimates.** Results use six episodes per run. The reported relative residual is $\|Ju_i - \theta_i u_i\|/|\theta_i|$. Normality is not established, so it is not an eigenvalue-error bound. The leading magnitude is stable, while the third-to-first ratio changes sharply with m; no multiplicity is inferred (App. O).

Arm	m	$\lvert\theta_1\rvert$	rel. residual	$\lvert\lvert\theta_1\rvert - \hat{\rho}\rvert$	$\lvert\theta_3\rvert/\lvert\theta_1\rvert$
S3DIS seed 777	16	1.342527	3.1×10^{-6}	1.4×10^{-7}	0.730093
	24	1.342527	2.1×10^{-5}	1.0×10^{-7}	1.000000
	40	1.342527	3.0×10^{-5}	5.7×10^{-8}	1.000000
S3DIS seed 2026	16	1.342604	9.7×10^{-6}	1.3×10^{-7}	0.738171
	24	1.342604	7.9×10^{-5}	9.8×10^{-8}	1.000000
	40	1.342604	1.2×10^{-5}	8.7×10^{-8}	1.000000
ScanNet fold 1	16	1.345405	4.3×10^{-5}	1.4×10^{-7}	0.797480
	24	1.345405	2.7×10^{-5}	1.2×10^{-7}	0.999967
	40	1.345405	1.1×10^{-6}	1.0×10^{-7}	1.000000

The leading magnitude is stable across $m \in \{16, 24, 40\}$ on three confirmatory runs, but the number of nearly equal leading Ritz values is not. The third-to-first ratio is 0.730–0.797 at $m = 16$ and approximately one at larger m. A synthetic matrix with a known repeated eigenvalue shows the same dependence of the returned count on the Krylov dimension. This demonstrates that the count is not an identifier of algebraic multiplicity. The audit also lacks a reported basis-orthogonality error, full Arnoldi factorisation defect, block start, and direct residual for every claimed band member. Consequently the estimates are exploratory and no spectral multiplicity, operator degeneracy, or model-family mechanism is inferred from them.

What survives is a consistency check on the one quantity the main text does use. $|\theta_1|$ agrees with $\hat{\rho}$ to within 8.7×10^{-8} on every run, which is what a dominant growth factor of $\hat{\rho}$ predicts, and the trajectory fit of Section 4.2 obtains that factor and the orthogonal factor μ without Arnoldi at all. Where the two disagree in kind is instructive. Arnoldi returns a band of near-equal magnitudes whose width depends on m, while the trajectory fit returns $\mu \approx 1.002$, distinct from $\hat{\rho} \approx 1.34$. The trajectory measurement is the one we report. It does not depend on resolving the band.

P Exploratory cross-run prediction study

An exploratory analysis asks whether measurements from depths 1–16 predict cumulative label changes or task change after depth 16. The task target uses performance at depth 16 as its baseline and performance at depth 64 as its endpoint. All predictor features stop at depth 16. Each outer leave-one-training-run-out fold selects one scalar feature inside an inner leave-one-out loop and evaluates the held-out run once.

Table A14: **The cross-run prediction study is inconclusive after leakage correction.** Outer leave-one-training-run-out R^2 from predictors at depths 1–16 to changes after depth 16 through depth 64, with the single feature for each family selected by inner leave-one-out only. There are 10 independent runs. "without DT" repeats the complete nested procedure after removing the sole run from the second family, which is also the most extreme target. Removing the sole external-family run changes R^2 from 0.099 to -0.938 for raw-residual flip prediction, from 0.991 to -0.161 for quotient-based task-drop prediction, and from 0.502 to 0.162 for output-based flip prediction; negative R^2 is worse than predicting the mean.

Family	flips	flips without DT	task drop	task without DT	n
raw	0.099	-0.938	-0.063	-1.526	10
quotient	-0.095	-1.806	0.991	-0.161	10
output	0.502	0.162	-1.180	-13.695	10
JVP	-0.255	-4.357	-2.716	-1.225	6

The population is not balanced. Nine runs are 3D refiners, one is the pretrained prefix-sums checkpoint, and only six runs have complete JVP measurements. The apparent positive quotient score for task change disappears when the sole external-family run is removed. Nested selection prevents one form of leakage, but cannot remove family/task confounding or make these ten runs exchangeable. The study therefore supplies no diagnostic ranking and remains supplementary.

Q REPRODUCTION AND IMPLEMENTATION CHECKS

LoopFormer adapter and provenance. The reference 1-block and 2-block configurations do not encode the recurrent block count in the field consumed by the local loader. The adapter therefore cross-checks the checkpoint name against the indexed block keys in the `safetensors` files and sets that count before construction. Under the recorded Transformers version, tied output weights were not restored automatically, so the adapter explicitly ties the language head to the input embedding, matching the reference architecture. These repairs and discovered block keys are logged. The post-audit lock pins each checkpoint to an immutable Hugging Face commit and records SHA-256 digests and byte counts for every weight, configuration, tokenizer, and remote-code file used by the adapter.

Environment. The runs use Ubuntu 22.04, CUDA 12.4, Python 3.10, PyTorch 2.5.1+cu124, and one RTX 4090 for the audit. Important pins include `numpy==1.26.4`, `torch-geometric==2.6.1`, and `torch_points3d==1.3.0` installed with `-no-deps`.

Retraining provenance, and what it does and does not license. The three confirmatory checkpoints and the contractive control behind Sections 4.3–4.5 are retrained versions. The five exploratory mechanism arms reuse their existing checkpoints. The original machine was decommissioned before the mechanism round, so the four new arms were trained from scratch on a rebuilt environment. Retrained runs enter the manuscript only through an acceptance gate applied before any mechanism number is computed, which compares the new checkpoint's $\hat{\rho}$, r_{late}, κ, and μ against the previously recorded values at tolerances fixed in advance (0.05 absolute on $\hat{\rho}$ and κ, 0.02 on μ, 15% relative on r_{late}). All four passed. The three confirmatory seeds reproduce $\hat{\rho}$ to 1.0×10^{-3} and κ to 6×10^{-4}; the contractive control reproduces $\kappa = 0.869$ against 0.867. The gate carries one special case: on a contracting arm the 15% relative band on r_{late} is an absolute band of 1.1×10^{-6} around a convergence floor, which would reject a correct reproduction, so contracting arms are checked against a floor-relative criterion plus a hard $\mu < 1$ instead. Whether an arm is contracting is read from the previously recorded value.

The gate is on *geometry only*, deliberately. Task-level landmarks do not reproduce and cannot be made to: the episode cache is built from a directory listing whose order is not sorted, and cache construction consumes the RNG, so a cache regenerated on a different filesystem draws the same class pairs from different scenes with different point counts. Peak, settle, and freeze steps move by several iterations (peak $7 \to 8$, settle $18 \to 22$, freeze $48 \to 39$) between the two caches. Geometric quantities are averaged over 1.6 million elements per episode and reproduce within the stated acceptance tolerances; task numbers from the two caches are never compared with each other, and no such comparison is made in this paper.

The auditor's non-interference claim, corrected. An earlier version of this check required the logits to be bitwise identical with the auditor attached and detached. That criterion is not satisfiable on this model, and the reason is not the auditor. Two forwards with the auditor *detached* differ by 1.0 in logit L_2 and disagree on 2.8% of point labels, for two independent reasons: the forward pass updates a prototype buffer by EMA without guarding on `self.training`, so consecutive forwards differ by construction; and the neighbourhood kernels accumulate with `atomicAdd`, whose order varies per launch and which `cudnn.deterministic` does not reach. The check now asks the answerable question: whether the auditor's effect is within the model's own run-to-run spread and whether the geometric quantities reproduce to $< 10^{-6}$. Measured: 0.65, 0.88 against a spread of 1.0, and 5.2×10^{-11}. Terminal r and $\hat{\rho}$ reproduce across forwards to 3×10^{-11}, because they average the nondeterminism away.

The check was repeated on a second machine for the runs of Section 4.5, and it replicates: the model's own spread was 1.18 and the auditor's effect 1.43, so 1.22 times the spread. The geometric

half of the criterion needed one correction, and the correction is a measurement. Terminal geometry reproduces across two audited forwards to 2.6×10^{-10} on the expansive control, and to 4.5×10^{-4} on a projective arm — six orders apart on the same machine, the same code and the same episode. The whole difference sits in r; $\hat{\rho}$, pinned by the rescale, reproduces to 1.8×10^{-7}. This contrast is consistent with different amplification of the host model's intrinsic nondeterminism; the source of that difference remains unidentified. A fixed absolute band nevertheless mis-applies to the contracting arms, exactly as it does for r_{late} in the retraining gate above, so the criterion is 10^{-6} absolute or 25% relative. Per episode the projective figure is about 5% of that arm's own r_{late}; the reported quantities are 100-episode means.

TF32. TF32 is disabled explicitly for both matmul and cuDNN in every audited run and the setting is recorded for every run. This is not a formality: TF32 keeps 10 explicit mantissa bits against `float32`'s 23, which raises the resolvable floor on q by roughly three orders and would place the entire non-shared measurement inside the rounding noise.

Checks that prevent silent measurement errors. The readout is captured from the model's forward pass, because reconstructing it outside the model can audit a different function. Finiteness guards use `isfinite`, not only `isnan`, because geometric growth reaches infinity first. Directional metrics are computed before overwriting the previous state. A synthetic radial trajectory $H_{t+1} = \rho H_t$ verifies $r_t = |\rho - 1|$ and $q_t = 0$. Rates and floors are separated mechanically rather than by eye. A decay rate is fitted only where the fitted window begins at least two decades above the floor it decays into. Otherwise, the rate is reported as absent. This is why Table 1 gives no μ for prefix-sums extrapolation. The three-way implementation includes synthetic pure-component cases plus reconstruction, pairwise-orthogonality, and energy-identity tests. Over the common resolvable window, the worst observed reconstruction error is 7.6×10^{-17}, the energy-identity error is 1.0×10^{-13}, and the normalised pairwise inner product is 6.8×10^{-11}; later floor-limited values are excluded from component attribution.