SEQUENTIAL UPDATES DO MORE THAN REMEMBER: COMPUTATION OVER HISTORY IN TEST-TIME LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Test-time learning is usually viewed as writing information into a fast state. We show that sequential test-time updates can do more: because each update acts on a state transformed by earlier updates, their dynamics can compute jointly and order-sensitively over history. We trace this historical computation through four stages: generation, visibility, behavioral use, and acquisition. For a broad class of smooth sequential updates with η-scaled history dependence, an interaction involving r distinct histories cannot appear before order η^r, and its leading order-sensitive structure is determined by update chronology. In an exactly solvable DeepLinear specialization, these interactions survive in the predictor but attenuate rapidly with support; nuisance-cancelling trajectory contrasts expose rather than amplify them. The same chronology-sensitive structure persists in nonlinear learners and across two sizes of the official pretrained Test-Time Training MLP (TTT-MLP). Task supervision then recruits causal pairwise composition through the evolving fast state in pretrained TTT-MLP; separately, ControlledTTT reveals functional contributions at support two, three, and four. Finally, dense training trajectories show that deeper historical support reaches stable behavioral usefulness later, consistent with a conditional small-step timescale mechanism. Together, these results identify sequential test-time adaptation as a mechanism not only for storing history, but for computing over it.

1 INTRODUCTION

Test-time learning and fast-weight systems modify an internal state while a sequence is processed (Sun et al., 2025; Schlag et al., 2021; Irie et al., 2021). These updates are usually interpreted as memory: each example writes information that later queries retrieve. But every update acts on a state already transformed by earlier examples. Earlier history can therefore change not only what the state contains, but how later history is processed. We ask whether this state dependence creates genuinely joint, order-sensitive *computation over history*. Here this means joint, chronology-sensitive effects created as earlier histories alter the updates applied to later ones, rather than contributions from independently computed updates combined only downstream.

Three histories give the running intuition. The first changes the state on which the second acts, and their joint effect changes how the third is processed. Reordering the same histories can then change the final state when update maps do not commute. This case cleanly exposes nested recomputation beyond a single pairwise interaction while remaining exactly analyzable; our structural results extend to arbitrary finite historical support.

Generation is only the first question. We next ask whether historical computation survives in the realized predictor and can be exposed at a useful scale, whether learners can use it for behavior, and when deeper historical contributions become useful during training. We organize the paper around these four stages:

$$\boxed{\text{generation} \longrightarrow \text{visibility} \longrightarrow \text{behavioral use} \longrightarrow \text{acquisition.}}$$

Historical support—the number of distinct histories jointly affecting a response—provides a common axis across them; chronology determines how those histories interact.

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.

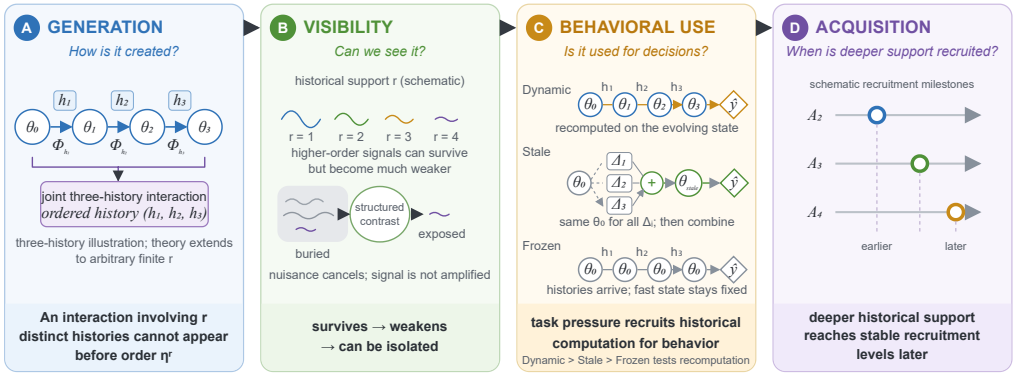


Figure 1: **The lifecycle of computation created by sequential updates.** State-dependent updates generate ordered multi-history interactions. We trace whether they survive and can be exposed, are recruited for behavior, and reach stable usefulness during training. Schematic quantities are marked as such.

For smooth sequential updates with η-scaled history dependence, an r-history interaction cannot appear before $O(\eta^r)$; chronology determines its first eligible order-sensitive structure. DeepLinear provides an exact microscope: the interaction survives in the realized predictor, but higher-support signals attenuate by more than four orders of magnitude, while nuisance-cancelling contrasts expose the buried structure. The chronology-sensitive structure remains detectable under nonlinear updates and across two sizes of the official pretrained Test-Time Training MLP (TTT-MLP).

After one-step grounding, an official pretrained TTT-MLP performs individual operations accurately yet shows a large two-step composition gap. Final-answer-only supervision recruits a fast state that the second operation can use, and interventions on that state causally alter subsequent computation. A separate controlled task extends behavioral use to three- and four-history support. Training trajectories then show a stable acquisition hierarchy: higher-support contributions become useful later, consistent with a conditional small-step mechanism.

The algebra of sequential composition is classical (Reutenauer, 1993; Butcher, 1972; Fliess, 1981); our contribution is the mechanistic pipeline that traces computation over history in TTT from arbitrary-support generation through function-level survival and visibility, causal behavioral recruitment and use, and acquisition.

Contributions. We provide (i) arbitrary-support theory for computation over history by smooth sequential updates; (ii) predictor survival, attenuation, and structural exposure, with nonlinear and pretrained persistence; (iii) causal task-supervised recruitment plus controlled behavioral use beyond pairwise interactions; and (iv) a stable-usefulness acquisition hierarchy with a conditional small-step timescale mechanism.

Table 1: **Order terminology.** Historical support is the common organizing axis throughout the paper; the remaining notions coincide only in specific constructions.

Notion	Meaning
Chronology	Sequence in which histories are processed
Historical support r	Number of distinct histories jointly affecting a response
Perturbative order	First nonzero power of η
Scalar value degree	Polynomial/Walsh degree in diagnostic values
Task depth	Number of composed task operations
Recruitment time	When a support-specific contribution becomes useful

2 Sequential Updates Generate Computation over History

Let history $i = (k_i, y_i)$ induce an update Φ_i^η of fast state θ. *Dynamic* processing applies $\theta_i = \Phi_i^\eta(\theta_{i-1}; y_i)$, recomputing each update on the evolving state. A *Stale* control evaluates each one-history delta from the common initial state and only then accumulates them; *Frozen* disables adaptation. Thus an earlier history can change the state on which a later Dynamic update is recomputed, and hence how that later history is processed.

With three histories, y_1 can change the state used to process y_2, and their joint effect can change how y_3 is processed.

Assume updates of the form

$$\Phi_i^\eta(\theta; y_i) = F_i^\eta(\theta) + \eta G_i^\eta(\theta; y_i),$$

where η is the step size, F_i^η is independent of y_i, and all y_i-dependence enters through the η-scaled term G_i^η.

Theorem 2.1 (Smooth historical-support generation)**.** *Consider a fixed finite sequence with distinct target sources whose updates have the form above. Suppose the required derivatives are uniformly bounded on a common compact trajectory neighborhood, and no other update, initialization, or readout depends on y_i. For a target-independent C^r observable h and distinct histories $i_1, \dots, i_r$,*

$$D_{y_{i_1}} \cdots D_{y_{i_r}} h(\theta_T) = O(\eta^r).$$

Every additional distinct history must enter through another η-scaled update, so a joint r-history response cannot appear below order η^r. The theorem excludes lower-order dependence; it does not require the first eligible coefficient to be nonzero. Proof and coefficient form appear in Section A.

From support to chronology. Historical support specifies which histories participate; chronology specifies how their updates compose. Reordering the same histories can change the first eligible interaction when the update maps do not commute. Near the identity,

$$\Phi_i^\eta = \mathrm{Id} + \eta X_i + O_{C^r}(\eta^2),$$

where X_i is the first-order update field. With pullback $U_i^\eta h = h \circ \Phi_i^\eta$ and Lie derivative $\mathcal{L}_X$, signed composition contrasts expose noncommutative chronology.

Theorem 2.2 (Connected chronology extraction)**.** *Under the regularity conditions of Appendix B, a multilinear Lie word ℓ satisfies*

$$\ell(U_1^\eta - I, \dots, U_r^\eta - I)h = \eta^r \mathcal{L}_{\ell(X_1,\dots,X_r)} h + O(\eta^{r+1}).$$

The signed trajectory contrast cancels every order below r, leaving the leading connected chronology-sensitive interaction. Specializing the learner, state, and readout may still make this eligible η^r coefficient vanish. Lie contrasts form a canonical connected subspace of the full proper-marginal cancellation space; the characterization is in Sections A and B.

Exact DeepLinear specialization. For $f(k; A, w) = w^\top A k$ with simultaneous squared-loss updates, the filtration becomes coefficientwise exact. Write $L_k(A, w) := (wk^\top, Ak)$ for the DeepLinear value-injection operator.

Proposition 2.3 (Exact DeepLinear support filtration)**.** *For every multi-index α,*

$$\mathrm{ord}_\eta(\theta_{t,\alpha}) \geq |\mathrm{supp}(\alpha)|.$$

If α is square-free with support $i_1 < \cdots < i_r$, then

$$[\eta^r v_{i_1} \cdots v_{i_r}]\theta_T = L_{k_{i_r}} \cdots L_{k_{i_1}} \theta_0.$$

At the first eligible order, the selected histories form a literal chronological chain. Static affine updates stop at support one; terminal combination of stale affine factors can create support two; Dynamic recomputation can continue the chain to arbitrary historical support. Full formulas are in Sections G and H.

Generation establishes that sequential state dependence can create joint, chronology-sensitive computation over history. The next question is whether this computation survives in the realized predictor and remains visible at a useful scale.

3 GENERATED COMPUTATION CAN BE HIDDEN

Section 2 establishes generation in the evolving state, not survival in the realized function: the readout can annihilate an interaction, and lower-support responses can overwhelm one that survives. In this section, we resolve these two obstacles exactly in DeepLinear, using it as an analytic microscope. Section 4 then asks which parts of this visibility picture persist under nonlinear mixing and pretraining.

3.1 SURVIVAL IN THE REPRESENTED PREDICTOR

For $f(k; A, w) = w^\top Ak$, many factor pairs represent the same $\beta = A^\top w$. Factor-only interactions may therefore be coordinate artifacts; survival must be tested in β.

Theorem 3.1 (Predictor survival and spherical strength)**.** *Let B_r be the leading square-free r-history predictor coefficient and $H_0 = A_0^\top A_0 + \|w_0\|^2 I$. With independent unit-spherical keys, B_r is nonzero almost surely whenever $\beta_0 \neq 0$ for even r, or $H_0 \neq 0$ for odd r, and*

$$\mathbb{E}\|B_{r+2}\|^2 = \lambda_d \mathbb{E}\|B_r\|^2, \qquad \lambda_d = \frac{6}{d} + \frac{10}{d^2},$$

where λ_d is the spherical pairwise strength factor. An independent spherical query contributes a further mean-square factor $1/d$.

Thus the interaction survives the factor-to-function map. When $\lambda_d < 1$, as in our dimensional regime, pair geometry attenuates with support; η^r supplies separate small-step suppression. For anisotropic keys, the positive map $Q \mapsto \mathcal{T}_\Sigma(Q)$ replaces λ_d. Details are in Sections E and K to M: the predictor recurrence is in K.3, followed by the moment law and generic survival proof.

3.2 ARBITRARY-SUPPORT VISIBILITY AND STRUCTURAL EXPOSURE

Survival is not visibility. Even when an r-history interaction survives in the predictor, lower-support responses can hide it. With exactly r Rademacher-valued histories, Walsh orthogonality canonically decomposes queried-output energy by support. Let E_s^η be the support-s energy and $N_{<r}^\eta$ the total nonempty proper-support nuisance:

$$E_s^\eta = \sum_{|S|=s} |\widehat{f}_S^\eta|^2, \qquad N_{<r}^\eta = \sum_{s=1}^{r-1} E_s^\eta.$$

Corollary 3.2 (DeepLinear arbitrary-support visibility)**.** *In the iid unit-spherical setting of Theorem 3.1, fix $r \geq 2$ and let $\eta \to 0$, with independent spherical query, finite η-independent initialization, and $H_0 \neq 0$.*

$$\frac{\mathbb{E}N_{<r}^\eta}{\mathbb{E}E_r^\eta} = C_r \eta^{-2(r-1)}[1 + O(\eta)], \quad C_{2m+1} = \frac{2m+1}{\lambda_d^m}, \quad C_{2m} = \frac{2m\,\mathrm{tr}(H_0^2)}{d\|\beta_0\|^2 \lambda_d^m},$$

The even branch requires $\beta_0 \neq 0$; otherwise its leading full-support denominator vanishes.

Thus the visibility barrier scales as $\eta^{-2(r-1)}$, making deeper interactions progressively harder to resolve against proper-support background. Singleton terms asymptotically dominate the nuisance. Proof and remaining scope are in Section O.

$r = 3$ **specialization.** For $r = 3$, the hierarchy becomes especially simple: singleton support dominates $N_{<3}$ at leading order, giving

$$\frac{\mathbb{E}E_1}{\mathbb{E}E_3} = \frac{3}{\eta^4 \lambda_d}[1 + O(\eta)].$$

Smaller steps clean the expansion but hide the cubic more strongly. At the fixed setting used below, this ratio is approximately 3.3×10^4.

A fixed chronology contrast instead attacks the nuisance directly:

$$\Delta_A = f_{213} + f_{312} - f_{132} - f_{231}$$

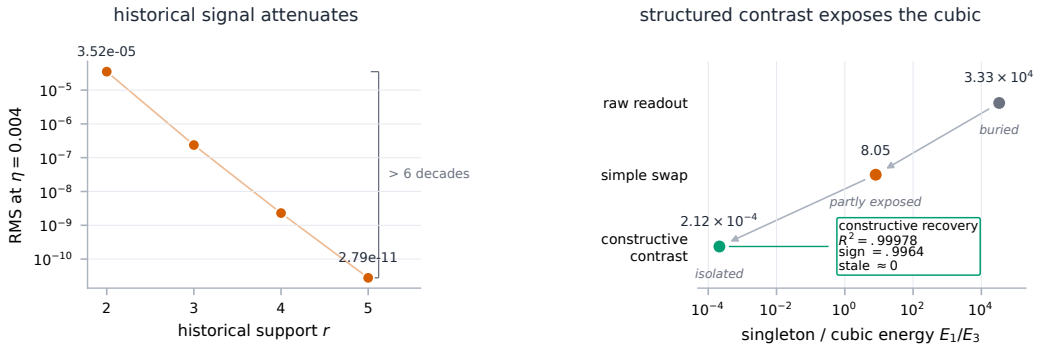


Figure 2: **Survival does not guarantee visibility.** *Left:* fixed-step historical-signal RMS illustrates support-wise attenuation in spherical DeepLinear, falling by more than six decades from support two to five. *Right:* a fixed structured contrast cancels lower-order nuisance and exposes the cubic interaction without amplifying it.

cancels the leading singleton while retaining a generically nonzero cubic. Figure 2 shows that a simple swap only partly helps, whereas the constructive contrast nearly isolates the cubic, with near-perfect held-out recovery and matched Stale at floor. The contrast does not enlarge the cubic interaction; it removes what hides it. Complete metrics and finite-step details are in Sections N and O.

More generally, the cubic construction illustrates a broader exposure principle:

Proposition 3.3 (Target-affine leading decontamination)**.** *Under Theorem 2.2, let scalar-target fields have the target-affine form $X_i = D_i + v_i L_i$. Then the proper-support-corrected contrast satisfies*

$$\Delta^{\eta}_{\ell,\mathrm{corr}} = \eta^r \Big(\prod_{i=1}^{r} v_i\Big) \mathcal{L}_{\ell(L_1,\dots,L_r)} h + O(\eta^{r+1}).$$

The vector-target result is the corresponding r-multilinear tensor.

Dividing by η^r recovers the nondegenerate full-support term up to $O(\eta)$; details are in Section C.

DeepLinear therefore resolves predictor survival, the arbitrary-support visibility hierarchy, and structural exposure exactly, but its clean analytic coordinates are special. Section 4 asks what persists under nonlinear mixing and pretraining.

4 The Structure Survives beyond the Analytic Microscope

DeepLinear resolves the mechanism exactly, but realistic learners deform its coordinates. We therefore ask whether chronology remains detectable under nonlinearity and pretraining, rather than whether every exact formula transfers.

Nonlinear updates. For $f(k; A, w) = w^\top \mathrm{GELU}(Ak)$, the histories still compose through non-commuting update fields, while the nonlinear readout mixes scalar value degrees, so a chronology contrast contains predictable leakage. At $r = 3$, a correction fixed from local derivatives reduces the low/cubic energy ratio from 2.13 to 3.2×10^{-4}, while the corrected response correlates .99984 with the true cubic component, without training a decoder. A locked $r = 4$ audit likewise reduces contamination by nearly $7000\times$, with corrected/true correlation .99148. Nonlinearity changes the nuisance that must be removed, not the underlying chronology.

Pretrained TTT-MLP. At fast-state and branch representations of the official 125M checkpoint (Sun et al., 2025), Dynamic trajectories retain support-two and support-three interactions and forward/reverse chronology while Stale response remains near numerical floor. Learned representations no longer preserve the privileged DeepLinear/GELU coordinates, so we fit a two-dimensional trajectory subspace calibrated only from mechanism trajectories, without downstream task labels, to

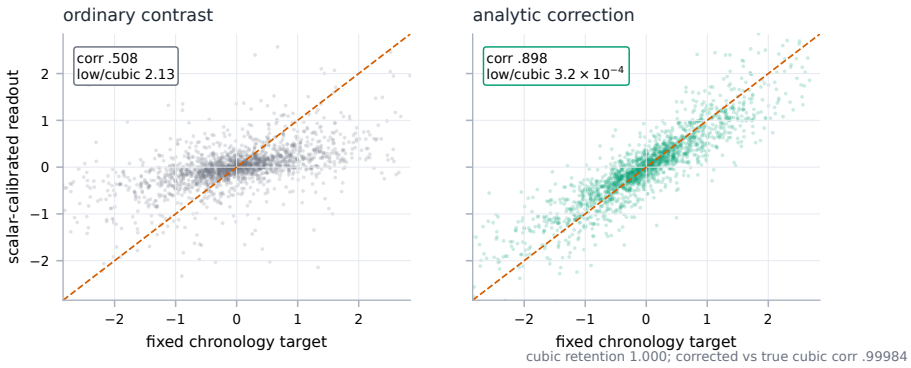


Figure 3: **Chronology survives nonlinear mixing.** In the controlled GELU setting, the ordinary readout mixes the $r = 3$ chronology target with lower-support leakage (left). An analytic correction removes this leakage and isolates the cubic component (right), without fitting a decoder or amplifying the signal.

separate cubic chronology response from lower-degree nuisance. On held-out data its low/cubic contamination ratio is approximately 3494, $13.1\times$ lower than the matched-random median of 45784, with all 32 batches beating that median. A separately fixed small-step audit of the same controlled one-token layer map gives support-two/three slopes $1.99/2.87$, consistent with the predicted scaling. A separately locked 350M cross-size replication under the same architecture-only midpoint rule reproduces support-two/three interactions and chronology in 5/5 fixed seeds; Stale stays near floor and fixed small-step slopes are $1.99/2.91$ (Section R.3).

These results establish detectable, theory-guided chronology structure in pretrained representations; stronger claims about a fixed global causal direction or untouched pretrained language behavior are outside the present evidence. Architecture taps, random controls, and causal boundaries are in Sections R to T.

Detectability alone does not establish behavioral use. Section 5 tests whether task pressure recruits this computation through supervision and interventions on the evolving fast state.

5 Historical Computation Can Be Recruited for Behavior

We first ask how task supervision converts local competence into causal two-step computation, then whether that behavior depends specifically on historical recomputation. A separate controlled experiment extends behavioral use beyond pairwise support.

5.1 Local competence and compositional use are distinct

We write D1, D2, and D3 for composition depths one, two, and three; for operations p_1, p_2, D2 requires predicting $p_2(p_1(s_0))$ while $s_1 = p_1(s_0)$ remains latent. We ground the official pretrained TTT-MLP on one-step permutations of three-symbol sequences and evaluate on held-out diverse symbols. Matched one-step accuracy reaches 96–98%, and explicitly providing the correct intermediate state restores this competence for a subsequent operation. Yet when that state remains latent, matched two-step accuracy is 22.2%, near six-way chance (16.7%). Local competence is therefore distinct from leaving a fast state that the next operation can use.

5.2 Final-answer supervision recruits causal fast-state use

Starting from the same grounded checkpoint, we fine-tune the slow model parameters using only final-answer supervision, while inference executes the two-step sequence through the fast state. The prompt never exposes the intermediate state s_1, and training includes no D1, D3, auxiliary, or representation loss. Seven of eight runs meet the predeclared gate (D1 $\geq 90\%$, D2 $\geq 80\%$, nontrivial D2 $\geq 75\%$); the eighth misses only the D1-retention threshold. D3 is never trained. Table 2

reports D3 at the earliest recruited checkpoint R^*; at the final checkpoints across all runs, it reaches $40.8 \pm 2.7\%$, indicating partial deeper generalization rather than solved D3.

D2 accuracy alone could reflect a depth-specific mapping rather than use of the intermediate fast state. We therefore intervene after the first operation and before the second. If that state is required, resetting it should damage D2; the observed drop is 67.1 ± 9.4 percentage points across recruited models. More strongly, transplanting a same-symbol donor state increases donor-follow behavior by 81.6 ± 11.7 percentage points, showing that state content redirects the subsequent computation toward donor history. States produced by different histories with the same s_1 are substantially more interchangeable than those with different s_1, distinguishing content-specific routing from generic cache damage.

This causally useful fast state is not a simple global code: a fixed linear six-way s_1 probe remains near chance, while the first action is almost perfectly decodable and global same-s_1 clustering remains weak. It therefore carries a causally functional, history-mixed intermediate-state signal, not a clean history-independent register; causal use need not imply linear decodability.

Table 2: **Task pressure recruits fast-state computation and aligns it with history.**

A. Behavioral recruitment

Checkpoint	D1 accuracy	D2 accuracy	D3 zero-shot
Grounded R0	98.2	22.2	20.8
Earliest recruited R^* ($n = 7$)	94.7 ± 2.9	88.4 ± 7.0	33.6 ± 5.3

B. Causal fast-state interventions

Intervention	Effect	Interpretation
Reset after first operation	D2 -67.1 ± 9.4 pp	intermediate state required
Donor transplant	donor-follow $+81.6 \pm 11.7$ pp	state content redirects behavior

C. Historical recomputation

Test	D1 effect	D2 effect	History-specific effect
Dynamic–Stale	$+13.7 \pm 4.3$	$+64.6 \pm 6.2$	D2–D1 DiD $= +50.9 \pm 9.1$ pp
Support-2 ablation	—	-72.3 ± 6.5	positive cost in $7/7$ models

5.3 Recruitment aligns behavior with historical recomputation

The interventions above show that the fast state matters; Dynamic–Stale asks whether it matters through sequential state-dependent recomputation. All seven recruited models pass the predeclared historical-alignment criterion: the D2 gap is substantially larger than the D1 gap, and functionally ablating the support-two contribution causes a large accuracy drop in every model (Table 2). Successful composition therefore aligns with sequential recomputation and pairwise historical support, not merely generic adaptation.

Because D2 prompts are longer than D1, we treat the DiD as supporting rather than standalone evidence; matched same-chunk analyses and the independent reset, transplant, and support-two interventions provide convergent support (Section T.4).

5.4 Useful support extends beyond two histories

The pretrained experiment gives the cleanest causal evidence at pairwise support. To ask whether behavioral use extends deeper, a separate ControlledTTT experiment composes fixed affine transformations of $(x, y) \in \mathbb{Z}_5^2$ for task depths one through four. Operation strings are disjoint across train, development, and test; history is unavailable at readout, so only the fast state carries episode content. A separate GRU baseline near ceiling establishes task learnability.

Across eight independent runs, mean Dynamic accuracy is $99.91, 70.84, 66.29, 47.03\%$ at task depths one through four. Macro accuracy is 71.02% for Dynamic, 52.64% for Stale, and 20% for Frozen, with Dynamic $>$ Stale $>$ Frozen in every run. The learner remains depth-limited but generalizes to unseen ordered strings.

To isolate the contribution that requires a set of histories jointly, rather than effects explained by smaller subsets, we evaluate the correct-class margin $m(S)$ under every retained subset and define its functional contribution by Möbius inclusion–exclusion (Harsanyi, 1963; Kang et al., 2024):

$$I_S = \sum_{R \subseteq S} (-1)^{|S|-|R|} m(R).$$

This decomposes the finite input–output map; it neither intervenes on a unique internal activation nor measures perturbative η-order. Removing the aggregate $|S| = 2$ contribution harms task depths two through four; support-three and support-four contributions add positive behavioral value at their corresponding depths. In the prospectively locked three-seed detailed replication, removing support-three and support-four contributions costs 7.88 and 4.36 percentage points at depths three and four, respectively, with positive effects in every seed. Shortcut audits, seed-level effects, and full reconstruction ladders are in Section U.

Together, task pressure recruits causal pairwise composition in pretrained TTT-MLP, while a controlled learner uses historical support two, three, and four. Having isolated these contributions, we finally ask when they become useful during training.

6 Higher-Support Contributions Reach Stable Usefulness Later

The final ControlledTTT models use several historical supports, but do not reveal when those contributions become useful. We operationalize acquisition through stable recruitment time, tracking the same behavioral decomposition through a separate matched cohort of eight runs, each evaluated at 18 checkpoints over 6,000 training steps.

Let $A_r(t)$ be the accuracy contribution of support r at step t, and normalize within each run as $N_r(t) = A_r(t)/A_r(6000)$. Recruitment time $t_q(r)$ is the earliest crossing of fraction q that stays above threshold at the next saved checkpoint, normalizing each support by its own final usefulness.

Seven of eight runs satisfy $t_{50}(2) < t_{50}(3) < t_{50}(4)$, and all eight satisfy the same strict ordering at t_{75}. Aggregate times are $1500/2000/2750$ steps at t_{50} and $2000/2500/3500$ at t_{75} for support $2/3/4$. Six runs show strong local non-monotonicity: the result is a stable middle/late acquisition hierarchy, not smooth stagewise emergence.

Control for ordinary learning. The median halfway time of ordinary Dynamic macro accuracy is 1,000 steps, before the aggregate historical crossings. Depth-specific task accuracy also develops with depth, so this limited control rules out all measured quantities improving at the same relative stage, but not every depth-dependent optimization effect (Section U.6).

Support-specific functional-signal RMS and output-space covariance both satisfy $r = 2 > 3 > 4$ in all 144 run–checkpoint comparisons. Next-step AdamW gains show no stable cross-support ordering, so the bridge to the timescale picture is at the output level, not the optimizer level.

6.1 A conditional timescale mechanism

A frozen-mode calculation gives the following conditional small-step timescale picture:

$$\boxed{O(\eta^r) \text{ amplitude} \implies O(\eta^{2r}) \text{ covariance} \implies O(\eta^{-2r}) \text{ learning time.}}$$

Theorem 6.1 (Conditional historical-mode timescale)**.** *Let a frozen scalar mode satisfy $z_r^\eta = \eta^r g_r + O_{L^2}(\eta^{r+1})$, with $c_r = \mathbb{E}[g_r^2] > 0$. For orthogonal retained modes with an aligned noiseless target, population squared-loss GD with common step α has $1 - q$ coefficient-error contraction time*

$$t_q(r, \eta) = \frac{-\log(1-q)}{\alpha c_r} \eta^{-2r} [1 + O(\eta)].$$

For finitely many nondegenerate modes, sufficiently small η therefore gives later recruitment at higher historical support.

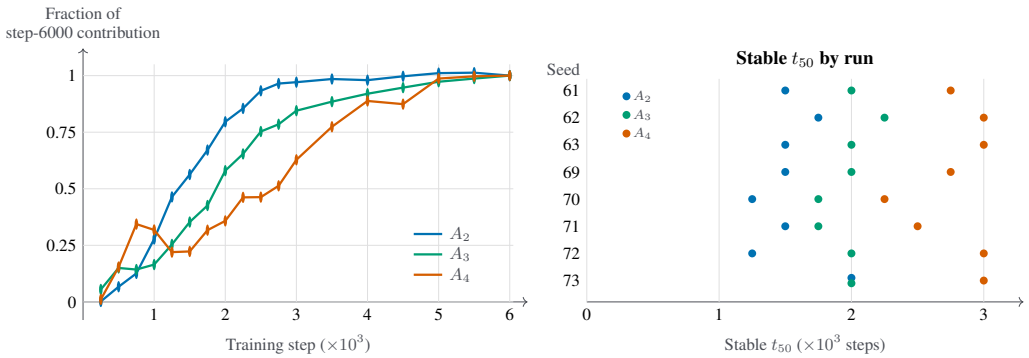


Figure 4: **Higher-support contributions reach stable usefulness later.** *Left:* aggregate contributions normalized at step 6,000, showing all checkpoints. *Right:* stable halfway times by run. Stable milestones progress from support two to three to four, while trajectories retain local non-monotonicity.

The mechanism is asymptotic and conditional: feature rescaling can reverse finite-η ordering, and output-level agreement does not establish an optimizer-level theory of end-to-end AdamW. DeepLinear supplies exact geometry factors; the proof, specialization, and finite-step counterexample are in Sections U.8–U.10.

7 Related Work

Test-time learning and fast weights. Fast weights, linear attention, and TTT place mutable state in sequence processing (Schmidhuber, 1992; Ba et al., 2016; Schlag et al., 2021; Irie et al., 2021; Sun et al., 2025). Related inference-time adaptation includes regression, KV binding, modular TTT, and Titans (Wang et al., 2025; Liu et al., 2026; Tang et al., 2026; Behrouz et al., 2025). Unlike prescribed higher-order attention (Zhang et al., 2025; Chakrabarti et al., 2026), we study history structure generated by sequential recomputation.

Sequential order effects. Continual task order and pairwise gradient noncommutativity study transfer (Li & Hiratani, 2025; Rukhovich et al., 2025; Sweeney, 2026); we trace arbitrary historical support through visibility, causal use, and acquisition.

Classical mathematical connections. Lie/BCH/Magnus, word-series, and Volterra machinery is classical (Magnus, 1954; Reutenauer, 1993; Butcher, 1972; Fliess, 1981; Murua & Sanz-Serna, 2017); online gradient descent is computationally expressive (Chatziafratis et al., 2019).

8 Discussion and Limitations

Sequential adaptation is more than memory: each update acts on a state shaped by earlier updates, changing how later history is processed and creating joint, chronology-sensitive effects.

Computational availability and behavioral use are distinct: an interaction may survive yet remain hidden, a detectable signal may remain unrecruited, and a causally important intermediate need not form a global linear register. Mechanistic evidence should not be reduced to the easiest coordinate or diagnostic.

The behavioral evidence remains controlled: task-supervised recruitment in pretrained TTT-MLP and higher-support value in ControlledTTT do not establish spontaneous use in untouched natural-language behavior. Acquisition is likewise scoped: deeper support tends to stabilize later, but trajectories are locally non-monotone and the frozen-mode analysis is conditional, not an end-to-end theory of AdamW dynamics.

Together, these results separate what sequential fast states can compute from what training teaches models to use.

AI USE STATEMENT

Generative AI tools were used in a supporting capacity for research ideation and execution, specifically for routine software assistance and symbolic and numerical cross-checks; for literature retrieval and discovery; and for manuscript structuring and language editing. The research questions, methodology, mathematical analysis, experimental design, and interpretation of results were determined by the authors. All AI-assisted material was reviewed by the authors, with mathematical and quantitative results independently verified against the corresponding derivations, proofs, stored evidence, and executable code, and cited primary sources manually inspected. The authors take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

Complete proofs and the experimental protocols underlying the main claims are provided in the appendix. Fixed seeds, checkpoint provenance, intervention definitions, and seed-level results are documented there. The anonymous supplementary data contains reviewer-facing result tables and compact protocol summaries.

REFERENCES

Jimmy Ba, Geoffrey E. Hinton, Volodymyr Mnih, Joel Z. Leibo, and Catalin Ionescu. Using fast weights to attend to the recent past. In *Advances in Neural Information Processing Systems*, volume 29, pp. 4331–4339, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/9f44e956e3a2b7b5598c625fcc802c36-Abstract.html.

Ali Behrouz, Peilin Zhong, and Vahab Mirrokni. Titans: Learning to memorize at test time. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3786. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/a4ca07aa108036f80cbb5b82285fd4b1-Abstract-Conference.html.

John C. Butcher. An algebraic theory of integration methods. *Mathematics of Computation*, 26(117):79–106, 1972. doi: 10.2307/2004720.

Sayak Chakrabarti, Toniann Pitassi, and Josh Alman. Poly-attention: A general scheme for higher-order self-attention. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/140edeced4facf41b14ca4f71a2322b7-Abstract-Conference.html.

Vaggos Chatziafratis, Tim Roughgarden, and Joshua R. Wang. On the computational power of online gradient descent. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 624–662, 2019. URL https://proceedings.mlr.press/v99/chatziafratis19a.html.

Stéphan Clémençon, Jérémie Jakubowicz, and Eric Sibony. Multiresolution analysis of incomplete rankings. *arXiv preprint arXiv:1403.1994*, 2014. URL https://arxiv.org/abs/1403.1994.

Michel Fliess. Fonctionnelles causales non linéaires et indéterminées non commutatives. *Bulletin de la Société Mathématique de France*, 109:3–40, 1981. doi: 10.24033/bsmf.1931.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 9574–9586, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/4f5c422f4d49a5a807eda27434231040-Abstract.html.

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah Goodman, and Christopher Potts. Inducing causal structure for interpretable neural networks. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 7324–7338, 2022. URL https://proceedings.mlr.press/v162/geiger22a.html.

Phil Hanlon and Patricia Hersh. A hodge decomposition for the complex of injective words. *Pacific Journal of Mathematics*, 214(1):109–125, 2004. doi: 10.2140/pjm.2004.214.109.

John C. Harsanyi. A simplified bargaining model for the n-person cooperative game. *International Economic Review*, 4(2):194–220, 1963. doi: 10.2307/2525487.

Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. Compositionality decomposed: How do neural networks generalise? *Journal of Artificial Intelligence Research*, 67:757–795, 2020. doi: 10.1613/jair.1.11674.

Kazuki Irie, Imanol Schlag, Róbert Csordás, and Jürgen Schmidhuber. Going beyond linear transformers with recurrent fast weight programmers. In *Advances in Neural Information Processing Systems*, volume 34, pp. 7703–7717, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/3f9e3767ef3b10a0de4c256d7ef9805d-Abstract.html.

Kazuki Irie, Imanol Schlag, Róbert Csordás, and Jürgen Schmidhuber. A modern self-referential weight matrix that learns to modify itself. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9660–9677, 2022. URL https://proceedings.mlr.press/v162/irie22b.html.

Justin Singh Kang, Yigit Efe Erginbas, Landon Butler, Ramtin Pedarsani, and Kannan Ramchandran. Learning to understand: Identifying interactions via the möbius transform. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1467. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/520b379123d16e41f85472e766846486-Abstract-Conference.html.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5156–5165, 2020. URL https://proceedings.mlr.press/v119/katharopoulos20a.html.

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, Dmitry Tsarkov, Xiao Wang, Marc van Zee, and Olivier Bousquet. Measuring compositional generalization: A comprehensive method on realistic data. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SygcCnNKwr.

Brenden Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2873–2882, 2018. URL https://proceedings.mlr.press/v80/lake18a.html.

Ziyan Li and Naoki Hiratani. Optimal task order for continual learning of multiple tasks. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 34578–34603, 2025. URL https://proceedings.mlr.press/v267/li25z.html.

Junchen Liu, Sven Elflein, Or Litany, Zan Gojcic, and Ruilong Li. Test-time training with KV binding is secretly linear attention. *arXiv preprint arXiv:2602.21204*, 2026. URL https://arxiv.org/abs/2602.21204. ICML 2026; arXiv version 4.

Wilhelm Magnus. On the exponential solution of differential equations for a linear operator. *Communications on Pure and Applied Mathematics*, 7(4):649–673, 1954. doi: 10.1002/cpa.3160070404.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-1262. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Tsendsuren Munkhdalai and Hong Yu. Meta networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2554–2563, 2017. URL https://proceedings.mlr.press/v70/munkhdalai17a.html.

Tsendsuren Munkhdalai, Alessandro Sordoni, Tong Wang, and Adam Trischler. Metalearned neural memory. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/182bd81ea25270b7d1c2fe8353d17fe6-Abstract.html.

Ander Murua and J. M. Sanz-Serna. Word series for dynamical systems and their numerical integrators. *Foundations of Computational Mathematics*, 17(3):675–712, 2017. doi: 10.1007/s10208-015-9295-3.

Mohammad Pezeshki, Amartya Mitra, Yoshua Bengio, and Guillaume Lajoie. Multi-scale feature learning dynamics: Insights for double descent. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 17669–17690, 2022. URL https://proceedings.mlr.press/v162/pezeshki22a.html.

Christophe Reutenauer. *Free Lie Algebras*. Oxford University Press, 1993. doi: 10.1093/oso/9780198536796.001.0001.

Alexey Rukhovich, Alexander Podolskiy, and Irina Piontkovskaya. Commute your domains: Trajectory optimality criterion for multi-domain learning. *arXiv preprint arXiv:2501.15556*, 2025. URL https://arxiv.org/abs/2501.15556.

Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1842–1850, 2016. URL https://proceedings.mlr.press/v48/santoro16.html.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations*, 2014. URL https://arxiv.org/abs/1312.6120.

Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. Linear transformers are secretly fast weight programmers. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9355–9366, 2021. URL https://proceedings.mlr.press/v139/schlag21a.html.

Jürgen Schmidhuber. Learning to control fast-weight memories: An alternative to dynamic recurrent networks. *Neural Computation*, 4(1):131–139, 1992. doi: 10.1162/neco.1992.4.1.131.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9229–9248, 2020. URL https://proceedings.mlr.press/v119/sun20b.html.

Yu Sun, Xinhao Li, Karan Dalal, Jiarui Xu, Arjun Vikram, Genghan Zhang, Yann Dubois, Xinlei Chen, Xiaolong Wang, Sanmi Koyejo, Tatsunori Hashimoto, and Carlos Guestrin. Learning to (Learn at test time): RNNs with expressive hidden states. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 57503–57522, 2025. URL https://proceedings.mlr.press/v267/sun25h.html.

John Sweeney. The geometry of sequential learning: Lie-bracket prediction of transfer order. *arXiv preprint arXiv:2606.24993*, 2026. URL https://arxiv.org/abs/2606.24993. ICML 2026.

Bohao Tang, Zhen Qin, Yuqi Pan, Zheng Li, Pengfei Liu, and Ya Zhang. Modular TTT: Rethinking test-time training as composable modules. *arXiv preprint arXiv:2608.07110*, 2026. URL https://arxiv.org/abs/2608.07110.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=uXl3bZLkr3c.

Ke Alexander Wang, Jiaxin Shi, and Emily B. Fox. Test-time regression: A unifying framework for designing sequence models with associative memory. *arXiv preprint arXiv:2501.12352*, 2025. URL https://arxiv.org/abs/2501.12352.

Yifan Zhang, Zhen Qin, and Quanquan Gu. Higher-order linear attention. *arXiv preprint arXiv:2510.27258*, 2025. URL https://arxiv.org/abs/2510.27258.

APPENDIX ROADMAP

The appendix follows the main scientific chain. Sections A–C give general support generation, complete cancellation, chronology, and leading exposure; Section D records the general-r constructions and validation. Sections E–M contain the geometry-dependent DeepLinear microscope, and Sections N–P give arbitrary-support visibility and nonlinear recovery. Section Q is the controlled nonlinear mechanism bridge. Section R, including its locked 350M cross-size replication in R.3, records official pretrained evidence, and Section S records the fixed-coordinate and adaptive-subspace scope boundary. Section T then separates controlled-language scope, grounding, D2-only recruitment, and historical alignment. Section U contains ControlledTTT behavioral use and the dense acquisition/timescale block. The remaining sections collect scoped limits, protocols, controls, and reproducibility details.

Protocol terminology. Throughout, *prospective* or *precommitted* denotes a specification recorded before the corresponding result-producing run; *locked* or *fixed* denotes choices held unchanged for the reported evaluation or replication. Analyses added afterward are identified as subsequent audits or robustness analyses.

A GENERAL SUPPORT GENERATION AND COMPLETE CANCELLATION

A.1 PROOF OF THE SMOOTH SUPPORT FILTRATION

Write the trajectory as a composition of the maps in Theorem 2.1. Differentiating with respect to a target y_i can first introduce that source only at update i, where

$$D_{y_i}\Phi_i^\eta = \eta D_{y_i} G_i^\eta.$$

Before and after this injection, state Jacobians and the higher state derivatives appearing in the multivariate chain rule are uniformly $O(1)$ on the common compact neighborhood. Every derivative tree contributing to $D_{y_{i_1}} \cdots D_{y_{i_r}} h(\theta_T)$ contains one such source derivative for each distinct selected memory. Hence every tree contains at least r explicit eta factors. The number of trees is finite at fixed horizon and derivative order, which proves the uniform $O(\eta^r)$ bound. Vector observables follow componentwise.

If the maps are formal power series, or analytic on a common polydisc, expansion in target monomials gives

$$\operatorname{ord}_\eta(\theta_{T,\alpha}) \geq |\operatorname{supp}(\alpha)|.$$

Analyticity is required for this convergent coefficient interpretation, not for the smooth mixed-response theorem. Repeated derivatives of one source do not buy additional eta powers; support counts distinct source memories.

A.2 UNIVERSAL PROPER-MARGINAL CANCELLATION

Let $W_i^\eta = U_i^\eta - I = O(\eta)$. For each $S = \{s_1 < \cdots < s_m\} \subseteq [r]$, expanding each factor but not its internal eta series gives the exact identity

$$\sum_{\pi \in S_r} c_\pi U_{\pi_1} \cdots U_{\pi_r} = \sum_{S \subseteq [r]} \sum_{\tau \in S_{|S|}} m_{S,\tau}(c) W_{s_{\tau(1)}} \cdots W_{s_{\tau(m)}}.$$

If every proper deletion marginal vanishes, only $S = [r]$ remains, and the contrast is $O(\eta^r)$ for every admissible near-identity family.

For necessity, attach independent formal scales t_i to smooth flows whose pullbacks begin $I + \eta t_i \mathcal{L}_{X_i} + O(\eta^2)$. The coefficient of $\eta^{|S|} \prod_{i \in S} t_i$ is

$$\sum_{\tau \in S_m} m_{S,\tau}(c) \mathcal{L}_{X_{s_{\tau(1)}}} \cdots \mathcal{L}_{X_{s_{\tau(m)}}}.$$

Ordered derivative words can be separated by smooth polynomial vector fields on a finite word-coordinate space, so universal cancellation forces every coefficient $m_{S,\tau}(c)$ to vanish. This proves

the universal-cancellation characterization summarized in Section 2 without assuming arbitrary linear operators are realizable learner updates.

The kernel is the classical top proper-marginal space in the injective-word complex, with dimension $!r$ (Clémençon et al., 2014; Hanlon & Hersh, 2004). The precise symmetric-group representation label may carry the conventional sign/dual twist; the kernel and dimension are the only facts used here.

A.3 Connected sector

For $c \in \mathcal{C}_r$, let $P_c(A_1, \dots, A_r) = \sum_\pi c_\pi A_{\pi_1} \cdots A_{\pi_r}$. Requiring P_c to be a derivation for arbitrary derivations A_i is equivalent to the primitive coproduct identity

$$\Delta P_c = P_c \otimes 1 + 1 \otimes P_c.$$

By the classical Friedrichs characterization, primitive tensor-algebra elements are exactly free-Lie elements (Reutenauer, 1993). Therefore the universally derivation-valued subspace of $\mathcal{C}_r$ is $\mathrm{Lie}(r)$. The contrast $[U_1, U_2][U_3, U_4]$ is the simplest disconnected witness outside that sector.

B General-r chronology theorem: conventions and proof

This appendix records the function-space version suppressed in the main text. Let $K \Subset K_+ \Subset \Omega \subset \mathbb{R}^n$, where K_+ contains every partial trajectory of length at most r from K for all sufficiently small $|\eta|$. Assume

$$\Phi_i^\eta = \mathrm{Id} + \eta X_i + \eta^2 R_i^\eta, \qquad X_i \in C^r, \qquad \sup_{|\eta| \le \eta_0} \|R_i^\eta\|_{C^r(K_+)} < \infty,$$

on a neighborhood of K_+, and let $h \in C^{r+1}(\Omega; \mathbb{R}^m)$.

On nested compact domains, Taylor's theorem and the chain rule give the scale estimates

$$\|(U_i - I)f\|_{C^{k-1}} \le C|\eta| \|f\|_{C^k}, \quad \|(U_i - I - \eta L_{X_i})f\|_{C^{k-2}} \le C\eta^2 \|f\|_{C^k}.$$

The common trajectory neighborhood makes the constants uniform through the fixed number of compositions.

Set $V_i = U_i - I$. A multilinear Lie polynomial vanishes when any leaf is replaced by the central identity. Expanding $U_i = I + V_i$ therefore gives the exact identity $\ell(U_1, \dots, U_r) = \ell(V_1, \dots, V_r)$. Put $A_i = \eta L_{X_i}$ and use the multilinear telescoping identity

$$\ell(V_1, \dots, V_r) - \ell(A_1, \dots, A_r) = \sum_{j=1}^{r} \ell(A_1, \dots, A_{j-1}, V_j - A_j, V_{j+1}, \dots, V_r).$$

Each summand has exactly one second-order difference $V_j - A_j$ and $r-1$ first-order factors. The scale estimates above therefore lose at most $r+1$ derivatives and, for $h \in C^{r+1}$, give a uniform $C^0(K)$ remainder $O(\eta^{r+1})$. Finally, $[L_X, L_Y] = L_{[X,Y]}$ for the convention $[X, Y]f = X(Yf) - Y(Xf)$, and induction over the bracket tree completes Theorem 2.2.

The theorem needs neither invertibility nor exact flow maps. For vector-valued h, it applies componentwise. It guarantees order at least r; the leading coefficient may vanish after field specialization or readout.

The coefficient space is the classical multilinear component of the free Lie algebra. A basis is

$$[x_{\sigma(1)}, [x_{\sigma(2)}, \dots, [x_{\sigma(r-1)}, x_r] \dots]], \qquad \sigma \in S_{r-1},$$

so its dimension is $(r-1)!$ (Reutenauer, 1993). For $r \ge 4$ this S_r module is generally reducible.

C Target-affine decontamination and normalized recovery

Let each target space Y_i be finite dimensional and write the first-order field as $X_i = D_i + L_i[y_i]$, where $L_i(\theta) : Y_i \to T_\theta \Theta$ is linear. By multilinearity of a Lie word ℓ,

$$\Delta_\ell^\eta(y) = \eta^r \sum_{S \subseteq [r]} \mathcal{L}_{\ell(Z_1^S, \dots, Z_r^S)} h + O(\eta^{r+1}),$$

where $Z_i^S = L_i[y_i]$ for $i \in S$ and $Z_i^S = D_i$ otherwise. Each summand is multilinear in exactly the selected target sources. Subtracting every term with $S \subsetneq [r]$ leaves

$$\Delta^\eta_{\ell,\mathrm{corr}} = \eta^r T_r(y_1, \dots, y_r) + O(\eta^{r+1}), \quad T_r = \mathcal{L}_{\ell(L_1[y_1],\dots,L_r[y_r])}h.$$

This proves Proposition 3.3; a basis expands T_r into coordinates but is not required by the tensor statement. The same leading subset expansion holds for a general $c \in \mathcal{C}_r$, although outside $\mathrm{Lie}(r)$ the leading operator need not be a single vector field.

For a canonical loss

$$\mathcal{L}(z, y) = A(z) - \langle y, z\rangle + c(y), \qquad z = f_\theta(k),$$

the negative-gradient field is

$$X(\theta; y) = -J_f^\top \nabla A(f_\theta(k)) + J_f^\top y,$$

which is target-affine. This includes scalar/vector squared loss, canonical softmax cross-entropy in logit coordinates, Bernoulli cross-entropy in canonical logit coordinates, and fixed label smoothing. Target-dependent weights, temperature, clipping, or nonlinear optimizer preprocessing require a separate audit.

If $\|R_\eta\| \le C|\eta|^{r+1}$ uniformly in $\Delta^\eta_{\mathrm{corr}} = \eta^r T_r + R_\eta$, then

$$\|\eta^{-r}\Delta^\eta_{\mathrm{corr}} - T_r\| \le C|\eta|,$$

giving the normalized-recovery claim in Section 3.2. For random episodes, relative mean-square convergence requires an L^2 remainder and finite nonzero $\mathbb{E}\|T_r\|^2$; scalar correlation additionally requires positive finite variance. Sign recovery follows from convergence in probability plus $\Pr(T_r = 0) = 0$, while a rate needs a margin assumption. None of these statements ensures native finite-step SNR or downstream utility.

D General-r constructions and validation

D.1 Explicit recursive contrasts

With $[A, B] = AB - BA$ and printed words denoting chronological execution, the fixed $r = 4$ contrast is

$$4321 - 4312 - 4213 + 4123 - 3214 + 3124 + 2134 - 1234,$$

and the $r = 5$ contrast is

$$\begin{aligned} &54321 - 54312 - 54213 + 54123 - 53214 + 53124 + 52134 - 51234 \\ &\quad - 43215 + 43125 + 42135 - 41235 + 32145 - 31245 - 21345 + 12345. \end{aligned}$$

All words are distinct. If N_r is the number of words in this right-nested family, prefixing and suffixing the new letter produce disjoint sets, so $N_r = 2N_{r-1} = 2^{r-1}$. No global minimality follows.

D.2 DeepLinear coefficient and slope checks

The result-producing protocol was locked before evaluation. Exact rational jets give coefficients of $\eta^0, \dots, \eta^{r-1}$ equal to zero and a nonzero η^r coefficient for $r = 2, 3, 4, 5$. Float64 rollouts use the same eta grid for all orders and give:

r	paths	$\dim \mathrm{Lie}(r)$	fitted slope	RMS at $\eta = .004$
2	2	1	1.9906	3.52×10^{-5}
3	4	2	2.9712	2.37×10^{-7}
4	8	6	3.9340	2.30×10^{-9}
5	16	24	4.8983	2.79×10^{-11}

Matched stale contrasts are below 10^{-3} conservative floor units and are mathematically null because the stale state is permutation invariant.

D.3 Full Lie(4) and nonlinear order four

The classical six-dimensional $\mathrm{Lie}(4)$ projector in the 24-word permutation space is approximately 6.39×10^4 cleaner than fixed random six-dimensional zero-sum subspaces for the prespecified operator-order signal. This does not imply value purity: under generic initialization, full value-degree-four energy is only 0.0129% in the corresponding DeepLinear diagnostic.

The canonical controlled GELU learner was reused without tuning. Its fixed quartic contrast has slope 3.9953, RMS 1.656×10^{-10} at $\eta = .008$, signal 497–945 floor units, and stale magnitude 4.88×10^{-4} floor units. The value spectrum is dominated by degree three (61.16%), while degree four is 0.598%. At the fixed $\eta = .008$ audit point, the median raw lower-support/degree-four energy is 111.103, the median corrected energy is 0.01726, and the median of the five per-seed reduction ratios is $6996.59\times$, with median degree-four retention 1.0 and median corrected/true degree-four correlation 0.9914847. Thus the operator-order theorem passes in a nonlinear learner precisely where a value-purity interpretation fails.

E General iid key-geometry propagation

For independent iid keys x, y with raw second moment $\Sigma = \mathbb{E}[xx^\top]$, set

$$M_{x,y} = (x^\top y)I + xy^\top + 2yx^\top, \qquad \mathcal{T}_\Sigma(Q) = \mathbb{E}[M_{x,y} Q M_{x,y}^\top].$$

Direct expansion and independence give

$$\mathcal{T}_\Sigma(Q) = \mu Q + 3(Q\Sigma^2 + \Sigma^2 Q) + 4\Sigma Q \Sigma + 5\,\mathrm{tr}(Q\Sigma)\Sigma, \qquad \mu = \mathrm{tr}(\Sigma^2).$$

Only the raw second moment is required in this iid-pair calculation. Dependence within a pair introduces joint fourth moments.

For disjoint independent chronological pairs,

$$Q_{m+1} = \mathcal{T}_\Sigma(Q_m), \quad Q_0^{\mathrm{even}} = \beta_0 \beta_0^\top, \quad Q_0^{\mathrm{odd}} = H_0 \Sigma H_0^\top.$$

Then $\mathbb{E}\|B\|^2 = \mathrm{tr}\, Q_m$, and an independent query with raw second moment Γ has scalar energy $\mathrm{tr}(\Gamma Q_m)$. In the spherical case $\Sigma = I/d$,

$$\mathrm{tr}\, \mathcal{T}_{I/d}(Q) = \left(\frac{6}{d} + \frac{10}{d^2}\right) \mathrm{tr}\, Q.$$

This trace identity does not make the full operator scalar on every matrix. Anisotropic eigendirections can contract or grow; correlated histories generally require conditional or multi-time propagation rather than powers of one fixed map.

F Full Notation and Coefficient Definitions

Model. The factor state is $\theta_t = (A_t, w_t) \in \mathbb{R}^{h \times d} \times \mathbb{R}^h$, the effective predictor is $\beta_t = A_t^\top w_t$, and both factors in a simultaneous update are evaluated at θ_{t-1}. Initialization and keys are independent of the formal values and eta variable.

Polynomial expansion. For finite T, the DeepLinear recurrence is polynomial, so

$$\theta_T(\eta, v_1, \ldots, v_T) = \sum_{\alpha \in \mathbb{N}_0^T} \theta_{T,\alpha}(\eta) v^\alpha$$

has finitely many nonzero terms. Each $\theta_{T,\alpha}$ is a polynomial in η. Vector and matrix eta valuation is the least degree at which any component is nonzero.

Square-free coefficients. For $S = \{i_1 < \cdots < i_r\}$, write $v_S = \prod_{i \in S} v_i$ and

$$C_S = [\eta^r v_S]\, \theta_T = (C_S^A, C_S^w).$$

Operator composition is right-to-left: $L_{k_{i_1}}$ acts first. The predictor and query coefficients are

$$B_S = [\eta^r v_S]\, \beta_T, \qquad Y_S(q) = q^\top B_S.$$

The corresponding leading contribution to the actual predictor is $\eta^r B_S v_S$, and to the squared predictor norm is proportional to η^{2r}.

Norms. We use absolute value for scalars, $\|\cdot\|_2$ for vectors, $\|\cdot\|_F$ for matrices, and

$$\|(A, w)\|_\Theta^2 = \|A\|_F^2 + \|w\|_2^2$$

only for fixed-coordinate factor-state comparisons.

Assumption boundaries. The leading-chain theorem is for a memoryless causal update with first jet $I + \eta(D_k + vL_k)$, value-independent eta-regular initialization, linear L_k, and distinct square-free histories. Persistent unscaled optimizer state, such as a momentum buffer that receives a gradient without eta scaling, requires a separate filtration and can invalidate the later-carry argument for the augmented state. Repeated numerical keys or values do not alter formal identities, but setting two formal values equal changes the meaning of historical support.

F.1 Static affine closure proof

Let $M_t = I - \eta h_t h_t^\top$. The fixed-feature update is

$$u_t = M_t u_{t-1} + \eta h_t v_t.$$

All M_t are independent of historical values. Therefore

$$u_T = M_T \cdots M_1 u_0 + \eta \sum_{i=1}^{T} M_T \cdots M_{i+1} h_i v_i.$$

This is affine in $(v_1, \ldots, v_T)$, proving the static baseline. The argument also allows value-independent, time-varying step sizes.

G Proof of the Eta-Support Filtration

We prove the factor-state statement; the predictor and query statements follow by polynomial convolution.

Proof of Proposition 2.3. At $t = 0$, only the zero multi-index occurs, so the claim holds. Assume it holds at time $t - 1$. Expand

$$A = \sum_\alpha A_\alpha v^\alpha, \qquad w = \sum_\alpha w_\alpha v^\alpha.$$

For the DeepLinear drift, coefficient extraction gives finite cubic convolutions

$$[v^\alpha]\, D_k^A = -\sum_{b+c+d=\alpha} (w_b^\top A_c k) w_d k^\top, \tag{1}$$

$$[v^\alpha]\, D_k^w = -\sum_{b+c+d=\alpha} (w_b^\top A_c k) A_d k. \tag{2}$$

Carry. Copying a coefficient from θ_{t-1} preserves both support and eta valuation.

Drift. For a nonzero convolution term $b + c + d = \alpha$, the induction hypothesis lower-bounds its internal eta valuation by

$$|\operatorname{supp}(b)| + |\operatorname{supp}(c)| + |\operatorname{supp}(d)| \geq |\operatorname{supp}(\alpha)|,$$

because $\operatorname{supp}(\alpha)$ is the union of the three supports. The external drift factor η raises the lower bound to $|\operatorname{supp}(\alpha)| + 1$. Summation and cancellation cannot lower valuation.

Value injection. Before update t, no coefficient contains the new formal variable v_t. An injection coefficient therefore has $\alpha_t = 1$ and equals

$$\eta L_{k_t}\big(\theta_{t-1,\bar{\alpha}}\big),$$

where $\bar{\alpha}$ removes coordinate t. Its eta valuation is at least

$$1 + |\operatorname{supp}(\bar{\alpha})| = |\operatorname{supp}(\alpha)|.$$

These are all contributions, completing the induction.

For $\beta = A^\top w$, any coefficient convolution has two supports whose union is $\operatorname{supp}(\alpha)$. The sum of their eta lower bounds is at least the union size. A fixed query is linear and preserves the valuation. □

Repeated powers. If drift later creates v_i^2 or v_i^3, multiple convolution factors can share index i. The sum of support sizes is still at least the size of their union. No square-free assumption is used in this proof.

Counterexample boundaries. Value-dependent initialization such as $\theta_0 = v_1 v_2$ violates the result at eta order zero. Negative eta powers likewise violate regularity. A nonsmooth normalized-gradient map may not possess the coefficient expansion at singular states.

H Proof of the Leading Ordered Chain

Leading-chain clause of Proposition 2.3. For $S \subseteq \{1, \dots, t\}$, define

$$C_t(S) = \left[\eta^{|S|} v_S\right] \theta_t.$$

For the empty set, the eta-zero state is always θ_0, since every non-carry update contains an external eta.

Fix nonempty square-free S and inspect update t.

If $t \notin S$, carry contributes $C_{t-1}(S)$, injection contains v_t and cannot contribute, and every drift product producing v_S partitions S among its factors. By Proposition 2.3, the internal eta degree is at least $|S|$; the external drift eta raises it to $|S| + 1$. Hence

$$C_t(S) = C_{t-1}(S).$$

If $t \in S$, the formal variable v_t is new. Neither carry nor drift can contain it before update t. Only injection contributes, and linearity of L_{k_t} gives

$$C_t(S) = L_{k_t} C_{t-1}(S \setminus \{t\}).$$

Applying the second recurrence at selected times and the first at every gap and later time yields

$$C_T(S) = L_{k_{i_r}} \cdots L_{k_{i_1}} \theta_0.$$

□

Why drift never supplies a same-order counterterm. An r-support square-free term at eta degree r has exactly enough eta budget for the first entry of each selected historical variable. Any drift, selected-time correction, or unselected update consumes another eta and therefore occurs at degree at least $r + 1$.

Optimizer caveat. The proof assumes a memoryless state or, more generally, no persistent unscaled auxiliary variable that can replay a selected gradient at later times. Coordinate-wise schemes sharing the same first-order jet preserve the displayed leading coefficient; momentum-augmented dynamics need their own statement.

I Proof and Consequences of the Alternating Pairing Formula

Theorem I.1 (Alternating Pairing Formula)**.** *Let* $C_S = (C_S^A, C_S^w) = L_{x_r} \cdots L_{x_1}(A_0, w_0)$. *For* $r = 2m$, $m \geq 1$,

$$C_S^A = (A_0 x_1) \left[\prod_{j=1}^{m-1} (x_{2j}^\top x_{2j+1})\right] x_{2m}^\top,$$

$$C_S^w = w_0 \prod_{j=1}^{m} (x_{2j-1}^\top x_{2j}).$$

For $r = 2m + 1$, $m \geq 0$,

$$C_S^A = w_0 \left[\prod_{j=1}^{m} (x_{2j-1}^\top x_{2j})\right] x_{2m+1}^\top,$$

$$C_S^w = (A_0 x_1) \prod_{j=1}^{m} (x_{2j}^\top x_{2j+1}).$$

Empty products equal one.

Proof of Theorem I.1. Write

$$(A_{r+1}, w_{r+1}) = L_{x_{r+1}}(A_r, w_r) = (w_r x_{r+1}^\top, A_r x_{r+1}).$$

The formulas hold at $r = 1$:

$$A_1 = w_0 x_1^\top, \qquad w_1 = A_0 x_1.$$

Assume the odd formulas hold at $r = 2m + 1$. One more injection gives

$$\begin{aligned} A_{2m+2} &= w_{2m+1} x_{2m+2}^\top \\ &= (A_0 x_1) \left[\prod_{j=1}^{m} (x_{2j}^\top x_{2j+1}) \right] x_{2m+2}^\top, \\ w_{2m+2} &= A_{2m+1} x_{2m+2} \\ &= w_0 \left[\prod_{j=1}^{m} (x_{2j-1}^\top x_{2j}) \right] (x_{2m+1}^\top x_{2m+2}), \end{aligned}$$

which are the even formulas. Applying $L_{x_{2m+3}}$ to those even formulas yields the next odd formulas. Induction completes the proof. □

I.1 Cuts and write order

The displayed inner products are literal blockwise cut edges. The endpoint $A_0 x_1 = 0$ kills every block that inherits it; $w_0 = 0$ kills the complementary provenance. The full state coefficient vanishes only when both blocks vanish.

At order two, forward and reverse coefficients are

$$\begin{aligned} C_{\mathrm{f}} &= \big((A_0 x_1) x_2^\top, w_0 s\big), \\ C_{\mathrm{r}} &= \big((A_0 x_2) x_1^\top, w_0 s\big), \end{aligned} \qquad s = x_1^\top x_2.$$

For iid spherical keys, with $a = \|A_0\|_F^2$ and $s_0 = \|w_0\|_2^2$,

$$\mathbb{E}\,\langle C_{\mathrm{f}}, C_{\mathrm{r}} \rangle = \frac{a}{d^2} + \frac{s_0}{d}, \qquad \mathbb{E}\,\|C_{\mathrm{f}}\|_\Theta^2 = \frac{a + s_0}{d}.$$

The normalized second-moment alignment is $(a/d + s_0)/(a + s_0)$. Thus an A-dominated initialization can exhibit strong directional rotation under reversal even though exchangeability preserves the complete marginal norm distribution.

J Factor-Space Random-Key Law

Proof of the factor-state moment equations. For independent spherical x, y,

$$\mathbb{E}(x^\top y)^2 = \frac{1}{d}, \qquad \mathbb{E}\,\|A_0 x\|_2^2 = \mathrm{tr}\big(A_0^\top A_0 \mathbb{E}[x x^\top]\big) = \frac{\|A_0\|_F^2}{d}.$$

Every scalar contraction in Theorem I.1 uses a disjoint pair of keys. It is independent of every other displayed contraction and, where present, of the endpoint $A_0 x_1$. Unit endpoint keys contribute no norm factor.

For $r = 2m$,

$$\mathbb{E}\,\big\|C_{2m}^A\big\|_F^2 = \frac{a}{d^m}, \quad \mathbb{E}\,\|C_{2m}^w\|_2^2 = \frac{s_0}{d^m}.$$

For $r = 2m + 1$,

$$\mathbb{E}\,\big\|C_{2m+1}^A\big\|_F^2 = \frac{s_0}{d^m}, \quad \mathbb{E}\,\big\|C_{2m+1}^w\big\|_2^2 = \frac{a}{d^{m+1}}.$$

Adding block energies completes the factor-state moment calculation. □

J.1 Raw-second-moment extension

The sphere is not necessary for a factor-state identity. Let iid keys have finite raw second moment $M = \mathbb{E}[kk^\top]$, and define

$$\nu = \operatorname{tr} M, \qquad \mu = \operatorname{tr}(M^2), \qquad \gamma_A = \operatorname{tr}(A_0^\top A_0 M).$$

Independence gives $\mathbb{E}(k_i^\top k_j)^2 = \mu$, $\mathbb{E}\,\|k\|^2 = \nu$, and $\mathbb{E}\,\|A_0 k\|^2 = \gamma_A$. Hence

$$\mathbb{E}\left\|C_{2m}^A\right\|_F^2 = \gamma_A \mu^{m-1}\nu, \quad \mathbb{E}\,\|C_{2m}^w\|_2^2 = s_0\mu^m,$$
$$\mathbb{E}\left\|C_{2m+1}^A\right\|_F^2 = s_0\mu^m\nu, \quad \mathbb{E}\left\|C_{2m+1}^w\right\|_2^2 = \gamma_A\mu^m.$$

Correlated histories break this product factorization; anisotropic iid keys do not.

J.2 Multiplicative skew

For independent sphere keys, $X = (x^\top y)^2$ satisfies

$$\mathbb{E}X = \frac{1}{d}, \qquad \mathbb{E}X^2 = \frac{3}{d(d+2)}.$$

For p independent contraction squares,

$$\mathrm{CV}^2\left(\prod_{j=1}^{p} X_j\right) = \left(\frac{3d}{d+2}\right)^p - 1.$$

The support remains bounded. The formula explains slow empirical convergence of high-order RMS without invoking a heavy-tailed distribution.

K Effective-Predictor Dynamics and All-Order Pairing

K.1 Exact finite-eta recurrence

Let $p = \beta^\top k$, $e = v - p$, $G = A^\top A$, and $s = \|w\|_2^2$. Simultaneous GD gives

$$A^+ = A + \eta e w k^\top, \quad w^+ = w + \eta e A k.$$

Direct multiplication yields the closed predictor-side recurrence

$$\beta^+ = \beta + \eta e(G + sI)k + \eta^2 e^2 p\, k, \tag{3}$$
$$G^+ = G + \eta e(\beta k^\top + k\beta^\top) + \eta^2 e^2 s\, kk^\top, \tag{4}$$
$$s^+ = s + 2\eta e p + \eta^2 e^2 k^\top G k. \tag{5}$$

Thus (β, G, s) is closed and hidden-coordinate free, while β alone is not. The combination $H = G + sI$ is convenient but is not closed without separate G, s, because the eta-squared terms distinguish $s\,kk^\top$ and $(k^\top G k)I$.

K.2 Leading partition formula

For $R \subseteq S$, let $C_R = (C_R^A, C_R^w)$ be the ordered factor chain, with $C_\varnothing = (A_0, w_0)$. Polynomial convolution gives

$$B_S = \sum_{R \subseteq S} (C_R^A)^\top C_{S\setminus R}^w. \tag{6}$$

At eta degree $|S|$, the two factor supports must form a disjoint partition; otherwise the value monomial has a repeated exponent or different support. This identity explains why projecting only C_S is insufficient: products of lower-order factor coefficients contribute at the same eta order.

K.3 Pair extraction

Introduce formal amplitudes u, v for two injections at keys x, y:

$$A_1 = A + uwx^\top, \quad w_1 = w + uAx,$$
$$A_2 = A_1 + vw_1y^\top, \quad w_2 = w_1 + vA_1y.$$

The coefficient linear in the first injection is

$$[u]\, A_1^\top w_1 = A^\top Ax + \|w\|_2^2\, x = (G + sI)x. \tag{7}$$

Extracting the bilinear term after two injections gives

$$\begin{aligned}[uv]\, A_2^\top w_2 &= (x^\top y)\beta + x(\beta^\top y) + 2y(\beta^\top x) \\ &= \mathcal{M}_{x,y}\beta. \end{aligned} \tag{8}$$

Pair-recurrence clause of Theorem 3.1. At eta order r, selected histories contribute only their value injections, by the same eta-budget argument as Proposition 2.3. Pair chronological injections as $(x_1, x_2), (x_3, x_4), \ldots$. Repeated use of Equation (8) propagates the effective coefficient through one $\mathcal{M}$ per pair. At even order the initial seed is β_0. At odd order, leave x_1 unpaired; Equation (7) supplies H_0x_1, and later pairs propagate it. Unselected and later writes carry the leading coefficient unchanged. □

The pair recurrence collapses the 2^r-term partition identity Equation (6) to $O(r)$ matrix-free operations. It is an exact coefficient identity, not an approximation to the factor expansion.

K.4 Low-order cancellation

At order two, write

$$B_{12} = M_{12}\beta_0, \qquad M_{12} = (x_1^\top x_2)I + x_1x_2^\top + 2x_2x_1^\top.$$

Thus $\beta_0 = 0$ is sufficient for cancellation, while special key/β_0 nullspaces can also cancel. For unit keys with $\rho = x_1^\top x_2 \neq 0$,

$$\det M_{12} = 2\rho^{d-2}(3\rho^2 - 1),$$

with the zero-correlation case treated directly. At order three, the odd-order recurrence depends on H_0, so $\beta_0 = 0$ is neither sufficient nor necessary for cancellation. For example, at $d = 1, h = 2$, $A_0 = (1, 0)^\top$, $w_0 = (0, 1)^\top$, and unit keys, $\beta_0 = 0$, $H_0 = 2$, and $B_3 = 8$.

L Proof of the Functional Random-Key Law

Let x, y be independent uniform unit vectors and let z be fixed or independent of them. Decompose

$$\mathcal{M}_{x,y}z = a + b + c, \quad a = (x^\top y)z, \quad b = x(y^\top z), \quad c = 2y(x^\top z).$$

Isotropy gives

$$\mathbb{E}\,\|a\|^2 = \frac{\|z\|^2}{d}, \quad \mathbb{E}\,\|b\|^2 = \frac{\|z\|^2}{d}, \quad \mathbb{E}\,\|c\|^2 = \frac{4\,\|z\|^2}{d}.$$

Each scalar cross expectation reduces to

$$\mathbb{E}[(x^\top y)(z^\top x)(z^\top y)] = \frac{\|z\|^2}{d^2}.$$

The coefficients of the three doubled cross terms are $2, 4, 4$, summing to ten. Therefore

$$\mathbb{E}\,\|\mathcal{M}_{x,y}z\|^2 = \left(\frac{6}{d} + \frac{10}{d^2}\right)\|z\|^2 = \lambda_d\,\|z\|^2\,. \tag{9}$$

Strength-law clause of Theorem 3.1. For even order, the pair operators use disjoint independent key pairs. Iterating Equation (9) from $z = \beta_0$ gives

$$\mathbb{E}\,\|B_{2m}\|^2 = \|\beta_0\|^2\,\lambda_d^m.$$

For odd order, x_1 is independent of all later pairs and

$$\mathbb{E}\,\|H_0 x_1\|^2 = \operatorname{tr}\big(H_0^2\mathbb{E}[x_1 x_1^\top]\big) = \frac{\operatorname{tr}(H_0^2)}{d}.$$

Iterating m later pair moments proves the odd-order formula in Theorem 3.1. Finally, for an independent spherical query,

$$\mathbb{E}_q(q^\top B)^2 = B^\top\mathbb{E}[qq^\top]B = \frac{\|B\|^2}{d},$$

which proves the queried-output laws. □

Gauge interpretation. The even law depends only on $\|\beta_0\|$ and is unchanged by every refactorization with the same initial predictor. The odd law depends on H_0, which is invariant under hidden orthogonal changes but not general $GL(h)$ refactorizations. Both moments concern a vector coefficient in input/predictor space.

M Generic Survival, Parity, and Gauge

Generic-survival clause of Theorem 3.1. Every coordinate of B_S is polynomial in the selected keys by the pair recurrence in Theorem 3.1.

For even $r = 2m$ and $\beta_0 \neq 0$, take every selected key equal to $u = \beta_0/\|\beta_0\|$. Since

$$\mathcal{M}_{u,u}z = z + u(u^\top z) + 2u(u^\top z),$$

and the iterate remains parallel to u, every pair multiplies it by four. Thus $B_S = 4^m\beta_0 \neq 0$ at this witness.

For odd $r = 2m+1$ and nonzero $H_0 \succeq 0$, choose a unit u with $u^\top H_0 u > 0$. Taking all keys equal to u gives a vector whose component along u is $4^m u^\top H_0 u > 0$. Hence the coordinate polynomials are not all identically zero in either case. Their common zero set is a proper algebraic set and has measure zero under any law absolutely continuous with respect to Lebesgue or sphere surface measure.

If $\beta_0 = 0$, the even-order formula is structurally zero. Since

$$H_0 = A_0^\top A_0 + \|w_0\|^2 I,$$

$H_0 = 0$ if and only if $A_0 = 0$ and $w_0 = 0$, which is the trapped odd-order initialization. □

M.1 Gauge audit

The standard hidden-basis action $(A, w) \mapsto (RA, R^{-\top}w)$, $R \in GL(h)$, preserves β at a fixed state. It transforms

$$G \mapsto A^\top R^\top RA, \qquad s \mapsto w^\top R^{-1}R^{-\top}w.$$

The Euclidean gradient step commutes with this action only when $R^{-\top} = R$, i.e., $R \in O(h)$. Therefore:

- factor-state norms and directions are coordinate dependent;
- every B_S and $q^\top B_S$ is a functional coefficient for the specified run;
- even B_{2m} is identical across all initial factorizations sharing β_0;
- odd B_{2m+1} generally changes under nonorthogonal refactorization through H_0.

In overcomplete or degenerate factorizations, the fiber representing one β may contain configurations outside a single regular $GL(h)$ orbit. Stating the central result directly in β avoids assuming otherwise.

M.2 Scalar parity witness

For $d = h = 1$, all keys one, and scalar factors a_0, w_0,

$$B_r = \begin{cases} 2^r a_0 w_0, & r \text{ even}, \\ 2^{r-1}(a_0^2 + w_0^2), & r \text{ odd}. \end{cases}$$

This witness makes the structural even cancellation at $\beta_0 = 0$ and generic odd survival transparent.

N Exact $r = 3$ Lie-coordinate recovery

Use word order $(123, 132, 213, 231, 312, 321)$ and commutator $[A, B] = AB - BA$. Direct expansion gives

$$[1, [2, 3]] = 123 - 132 - 231 + 321 = (1, -1, 0, -1, 0, 1) = w_B,$$

and

$$[2, [1, 3]] = 213 - 231 - 132 + 312 = (0, -1, 1, -1, 1, 0) = w_A.$$

The vectors are independent and $\dim \mathrm{Lie}(3) = 2$, proving $\mathrm{span}\{w_A, w_B\} = \mathrm{Lie}(3)$ exactly under the manuscript's trajectory convention. Thus the standard-S_3 description remains correct, while its representation-theoretic identification is classical.

For the recursive family $C_2 = [2, 1]$, $C_r = [r, C_{r-1}]$, prefixing and suffixing the new letter produce disjoint word sets, hence $N_r = 2N_{r-1}$ and $N_r = 2^{r-1}$. This does not establish global minimality.

O Arbitrary-support visibility and the complete $r = 3$ construction

O.1 Walsh visibility hierarchy

Fix exactly $r \geq 2$ histories under the simultaneous DeepLinear update, with iid unit-spherical keys, an independent unit-spherical query, iid Rademacher values, and fixed finite η-independent initialization. Write

$$\beta_0 = A_0^\top w_0, \qquad H_0 = A_0^\top A_0 + \|w_0\|^2 I, \qquad f_r^\eta(q; v) = \sum_{S \subseteq [r]} \widehat{f}_S^\eta v_S.$$

The Walsh characters are orthonormal under the value-sign average, so the coefficients are unique and

$$E_s^\eta = \sum_{|S|=s} |\widehat{f}_S^\eta|^2, \qquad N_{<r}^\eta = \sum_{s=1}^{r-1} E_s^\eta$$

is the exact conditional energy of all nonempty proper-support components, with no cross terms. The visibility corollary concerns $\mathbb{E}_{k,q} N_{<r}^\eta / \mathbb{E}_{k,q} E_r^\eta$, not an expectation of episode-wise ratios.

For a fixed $|S| = s$, the support filtration and predictor recurrence give

$$\widehat{f}_S^\eta = \eta^s q^\top B_s(k_S) + O_{L^2}(\eta^{s+1}).$$

Set

$$a = \frac{\mathrm{tr}(H_0^2)}{d^2}, \qquad b = \frac{\|\beta_0\|^2}{d}.$$

Spherical key/query moments and the $\binom{r}{s}$ equal-law supports yield

$$\mathbb{E} E_s^\eta = \binom{r}{s} \eta^{2s} \begin{cases} b\lambda_d^{s/2}, & s \text{ even}, \\ a\lambda_d^{(s-1)/2}, & s \text{ odd}, \end{cases} + O(\eta^{2s+1}).$$

Hence, for fixed finite r and $H_0 \neq 0$,

$$\mathbb{E} N_{<r}^\eta = ra\eta^2 + O(\eta^3),$$

because every other nonempty proper support contributes $O(\eta^4)$ or smaller. Dividing by the parity-specific full-support energy proves Corollary 3.2. For odd $r = 2m + 1$ the factor a cancels, giving $(2m + 1)\lambda_d^{-m}$; for even $r = 2m$, a/b gives

$$\frac{2m \operatorname{tr}(H_0^2)}{d\|\beta_0\|^2 \lambda_d^m}.$$

The independent query factor $1/d$ appears in both numerator and denominator; the remaining d in the even constant comes from $\mathbb{E}\|H_0 k_1\|^2 = \operatorname{tr}(H_0^2)/d$.

The statement is pointwise as $\eta \to 0$ for fixed finite r, d, h and fixed initialization. Uniformity over an initialization family additionally requires bounded factor norms and a leading denominator bounded away from zero. If $\beta_0 = 0$, the leading even full-support coefficient vanishes and the displayed leading-order law is undefined; the exact finite-η energy may start later. If $H_0 = 0$, then $A_0 = w_0 = 0$ and the update is trapped. Finally, λ_d is a strength factor: it attenuates paired coefficients only when $\lambda_d < 1$ and amplifies them in sufficiently small dimension.

O.2 $r = 3$ SPECIALIZATION

For $r = 3$, the preceding expansion gives

$$\mathbb{E}E_1^\eta = 3\eta^2 \frac{\operatorname{tr}(H_0^2)}{d^2} + O(\eta^3), \qquad \mathbb{E}E_3^\eta = \eta^6 \lambda_d \frac{\operatorname{tr}(H_0^2)}{d^2} + O(\eta^7).$$

Moreover $\mathbb{E}E_2 = O(\eta^4)$, so total proper-support nuisance and singleton energy have the same leading term. This recovers the $r = 3$ specialization of Corollary 3.2 exactly, including its constant.

O.3 PERMUTATION KERNEL AND CUBIC IMAGE

In ordering $(123, 132, 213, 231, 312, 321)$, the normalized leading singleton map has kernel

$$\operatorname{span}\{(0, -1, 1, -1, 1, 0), (1, -1, 0, -1, 0, 1)\}.$$

Every vector in this kernel is zero sum. In the formal cubic basis $(\rho_{23}u_1, \rho_{13}u_2, \rho_{12}u_3)$, the two displayed basis vectors map to $(-1, 0, 1)$ and $(0, -1, 1)$, respectively, proving injectivity of the leading cubic map on the kernel. The finite-setting audit below uses $A_0 = \sqrt{\alpha} I_d$ and $w_0 = 0$, hence $H_0 = \alpha I_d$, with $\alpha = \eta^p$ and fixed $p > 1$. This eta-dependent initialization is specific to this joint construction and is distinct from the eta-independent initialization in Corollary 3.2. For w_A, the leading term is

$$\alpha\eta^3 q^\top \{\rho_{12}k_3 - \rho_{23}k_1\} v_1 v_2 v_3 = \alpha\eta^3 q^\top \mathcal{A}_{k_1,k_3} k_2 \, v_1 v_2 v_3,$$

where $\mathcal{A}_{x,y} z = y(x^\top z) - x(y^\top z)$.

At $d = 8, \eta = .1, p = 2$ (so $\alpha = .01$), held-out Walsh audits give $E_1/E_3 = 33320.6$ for raw output, 8.0478 for a simple swap, and 2.122×10^{-4} for w_A. The fixed analytic readout gives NMSE 2.18×10^{-4}, $R^2 = .999782$, correlation .999893, and sign accuracy .9964 on unseen episodes. No contrast weights or decoder were fit to those episodes.

P GELU CHRONOLOGY GEOMETRY AND DECONTAMINATION

For $f_x(A, w) = w^\top \phi(Ax)$ and squared-loss field $X_i(v_i) = (v_i - f_i)g_i$, $g_i = \nabla f_i$, exact permutation algebra cancels update words of length at most two in w_A. With $[X, Y] = DY[X] - DX[Y]$,

$$\Delta_A = \eta^3 \nabla f_q^\top [X_2, [X_1, X_3]] + O(\eta^4).$$

Trilinearity in the episode-constant values splits this leading polynomial into Walsh supports. The cubic is

$$\eta^3 v_1 v_2 v_3 \nabla f_q^\top [g_2, [g_1, g_3]],$$

and the fixed analytic correction subtracts the complementary support-< 3 part. It is an explicit ordinary-episode function of the initial state, keys, values, query, known GELU derivatives, and local fields; it uses no counterfactual value cube, target, fitted coefficient, or learned decoder.

For exact GELU, $\phi'(z) = \Phi(z) + z\varphi(z)$, $\phi''(z) = (2 - z^2)\varphi(z)$, and $\phi'''(z) = (z^3 - 4z)\varphi(z)$. The nested correction levels successively retain the Jacobian geometry, add ϕ'', and add ϕ'''. Median five-seed results are:

Table 3: **Controlled nonlinear bridge.** Slopes use two fixed small-step windows; values shown are medians over independent seeds.

Probe	Expected qualitative signature	Observed result
$r = 2$ small-η scaling	η^2	slope 2.00
$r = 3$ small-η scaling	η^3	slope 3.00
$r = 3$ dynamic/stale output	dynamic larger in aggregate	median $16.2\times$
$r = 2$ chronology	nonzero order effect	relative difference 0.33
$r = 3$ chronology	nonzero order effect	relative difference 0.74

readout	corr. to DL target	sign	R^2	low/cubic
ordinary w_A	.5079	.7253	.2576	2.1346
Jacobian correction	.8889	.8754	.7902	—
second derivative	.8981	.8789	.8066	—
full local correction	.8981	.8789	.8066	3.209×10^{-4}

The full correction retains measured cubic energy to reported precision and has correlation .99984 with the post-hoc true GELU cubic; the table instead reports correlation to the fixed DeepLinear target. The true cubic itself has about .90 correlation with the fixed DeepLinear target; thus decontamination exposes the available nonlinear chronology rather than forcing an exact DeepLinear identity.

Q Controlled Nonlinear Mechanism Bridge

Learner and extraction. The learner is $f(k; A, w) = w^\top \mathrm{GELU}(Ak)$, with both factors updated by one simultaneous squared-loss SGD step per memory. For selected histories S, $|S| = r$, values are $v_i = as_i$ with $s_i \in \{-1, +1\}$. For a fresh-query output F, the extracted full-support mixed difference is

$$c_r(a) = 2^{-r} \sum_{s \in \{-1,+1\}^r} \left(\prod_{i=1}^{r} s_i \right) F(as).$$

The same transform is applied to the flattened fast state. This finite-amplitude transform cancels any response omitting a selected history; it is not a claim that the nonlinear response is polynomial in the values.

Dynamic and stale controls. Dynamic writes evaluate gradient t at the state produced by histories $1, \ldots, t-1$. The stale control evaluates every memory gradient at the common initial state and sums those updates. Histories, initialization, query, loss, learning rate, amplitude, and number of writes are otherwise identical. The stale state is affine in the values and therefore has zero support-r coefficient for $r \geq 2$, although the terminal GELU predictor can mix that affine state at output.

Replication and robustness. Ten freshly executed replication seeds give lower-window median slopes 1.99998 and 2.99966 for support two and support three; a second fixed window gives 1.99994 and 2.99892. At support three, the paired dynamic/stale output ratio has median 16.24 and IQR 6.58–55.83; dynamic exceeds stale for nine of ten primary seeds and 29 of 30 total condition-seed pairs. The maximum stale-state support-three coefficient is 0.0054 conservative floor units. Four small perturbations of dimension and initialization scale preserve the fixed signal, slope, separation, chronology, amplitude, and null-control gates.

Numerical and chronology controls. Eta-zero and static-feature interactions are at the empirical floor. Normalized coefficients at amplitudes $a = 0.5$ and $a = 1$ agree to median relative discrepancy 4.83×10^{-7}. Stale forward/reverse discrepancies remain below 2.44×10^{-4} floor units, while every dynamic seed changes above floor. All 1,080 float64 cube runs are finite. The initial protocol was fixed before evaluation; subsequent replication seeds and robustness perturbations were added without changing the original success criteria.

Scope. This controlled nonlinear bridge is scoped to the predicted $r = 2, 3$ small-step scaling, sequential-recomputation dependence, and chronology. Exact DeepLinear pairing recurrences, the λ_d law, pretrained-language-model behavior, and performance claims lie outside this bridge.

R TTT-MLP Architecture and Pretrained Evidence

R.1 Released/default-initialization localization

We instrumented the official PyTorch implementation of TTT-MLP from Sun et al. (2025), repository `https://github.com/test-time-training/ttt-lm-pytorch`, pinned at commit `cd831db10c8c9a0f6340f02da5613316a8a92b67`. The evidence level is published-architecture layer-level TTT: the released layer was instantiated in float64 at its default initialization, without a pretrained checkpoint or language-model evaluation. Dynamic and stale updates use the same histories, initialization, step scale, and Walsh mixed-difference probe; stale writes evaluate all updates at the common initial fast state.

For a fresh query q, the official TTT-MLP response follows

$$Z_1 = qW_1 + b_1, \quad X_2 = \mathrm{GELU}(Z_1), \quad Z_2 = X_2W_2 + b_2, \quad R = \text{TTT-LN}(Z_2),$$

followed by the residual canonical response $XQW = q + R$, post-normalization, and output projection. We tapped the complete vector coefficient at every listed stage rather than selecting a scalar direction after observing the result.

Figure 5 and Table 4 localize the same fixed probe at the response and coefficient levels, respectively.

The exact stale null first becomes detectable after GELU, grows after the second factorized response, and becomes large enough to obscure clean dynamic-over-stale amplitude separation at TTT LayerNorm and the canonical output. In particular, the ratios 0.38 and 0.44 do *not* mean that 38 percent or 44 percent of the dynamic interaction survives. They mean that, at those taps, downstream nonlinear processing has produced a larger stale full-support coefficient, so amplitude separation is no longer a clean recomputation diagnostic. Dynamic history dependence has not disappeared: median forward/reverse relative differences for $r = 3$ are 0.78 in the fast state, 1.09 at canonical XQW, and 1.15 after the final projection, while stale chronology remains at floor.

At default initialization, the canonical output does not retain a clean dynamic/stale amplitude diagnostic, yet chronology remains exposed. This localizes a distinction between interaction generation in the fast state and its exposure through nonlinear readout. The probe characterizes one released layer; exact DeepLinear scaling, training effects, and language-model behavior are addressed separately or remain outside this experiment.

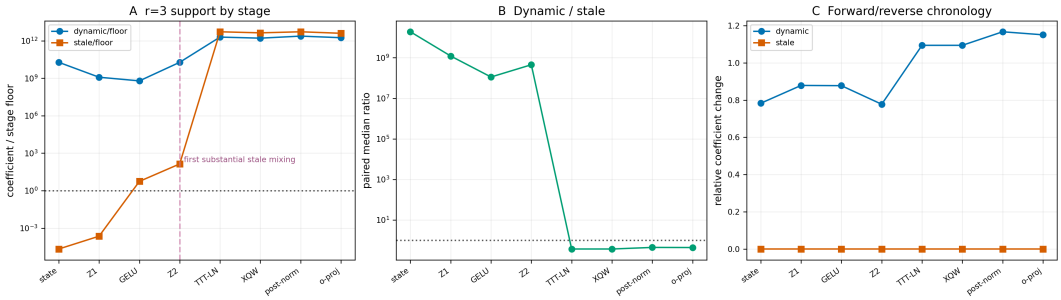


Figure 5: **Where generation and output exposure separate in released TTT-MLP.** Stale support-three interaction is at floor in the fast state and first affine query, appears after GELU, and grows through the nonlinear response and TTT LayerNorm. The stage-wise dynamic/stale amplitude comparison is not a retention fraction: chronology remains visible through canonical and final outputs even where stale mixing is larger. This probe uses released initialization, not a pretrained checkpoint.

Table 4: **Support-three readout localization at the fixed largest step.** Values are medians over five locked seeds. The Dyn./stale column compares coefficient magnitudes produced by two protocols at the same tap; it is not the fraction of a dynamic interaction retained through the network.

Stage	Dynamic	Stale	Dyn./stale amplitude	Dyn. slope	Interpretation
fast state	5.40×10^{-4}	6.00×10^{-19}	1.90×10^{10}	2.61	clean state
Z_1, pre-GELU	3.44×10^{-5}	6.38×10^{-18}	1.21×10^{9}	2.66	clean affine query
$\mathrm{GELU}(Z_1)$	1.76×10^{-5}	1.57×10^{-13}	1.12×10^{8}	2.66	stale term detectable
Z_2, pre-TTT-LN	5.40×10^{-4}	3.92×10^{-12}	4.51×10^{8}	2.61	stale mixing grows
canonical XQW	2.25×10^{-1}	5.99×10^{-1}	0.38	2.47	separation obscured
final projection	1.60×10^{-1}	3.60×10^{-1}	0.44	2.28	non-clean output

R.2 Locked pretrained-checkpoint bridge

We next probed one official pretrained checkpoint. Its locked identity is:

```
Test-Time-Training/ttt-mlp-125m-books-2k
revision 254be4aaff95c3421d7207540b9eb50b468bc0ef
  state SHA-256 bbbd3be437bb71047f839e16c658ecaf
                23e948d3461ff6fcf2d8e8df49ff1980
```

The official JAX implementation from https://github.com/test-time-training/ttt-lm-jax was pinned at commit `6f529b124c7fb5879b33c06926408b15add1d82f`. Before observing interaction results, the 12-layer architecture and predeclared midpoint rule selected zero-based layer $\lfloor 12/2 \rfloor = 6$. We used fixed seeds 700–704, with no checkpoint, layer, seed, probe, or eta-window search; there was no training or language-model evaluation.

The controlled interface sets $XQ_i = XK_i = k_i$ and $XV_i - XK_i = v_i u$. For each selected history set, the full-support vector coefficient is extracted by the same Boolean-cube/Walsh mixed difference. Dynamic writes apply the official sequential one-token update. The stale control sums the corresponding one-token deltas after evaluating each at the common initial fast state; histories, initialization, step scale, amplitudes, and query are otherwise shared. The conservative numerical floor is the maximum of the eta-zero coefficient, amplitude-zero coefficient, and $128\epsilon_{64}\max(1, \text{maximum raw tap norm})$.

Table 5 reports all predeclared taps, controls, and fixed-window slopes for the 125M checkpoint.

The locked recomputation and chronology signatures therefore persist after pretraining in this probe. Dynamic support-two and support-three fast-state interactions are far above floor, while both stale fast-state coefficients are below it. At support three, dynamic chronology is also far above floor in the fast state, canonical TTT response, and controlled layer-interface output, whereas stale chronology is below 0.003 times floor. Eta-zero, amplitude-zero, and secondary-amplitude controls pass, and all probe runs are finite.

Table 5: **Locked pretrained TTT-MLP probe at zero-based layer 6.** Values are medians over five locked seeds. "Floor ratio" is magnitude divided by the conservative numerical floor. Chronology rows report the norm of the forward/reverse coefficient difference. The controlled layer-interface output uses trained gating, post-normalization, and output projection, but is not a full-model natural-text output.

Probe	Support	Dynamic	Stale	Relative-to-floor / interpretation
Fast-state interaction	2	0.407	6.89×10^{-16}	dynamic 2.98×10^{11}; stale 4.75×10^{-4}
Fast-state interaction	3	0.0537	6.51×10^{-16}	dynamic 3.94×10^{10}; stale 4.04×10^{-4}
Fast-state chronology	3	0.0449	5.66×10^{-16}	dynamic 3.30×10^{10}; stale 3.49×10^{-4}
Canonical-response chronology	3	0.0650	7.91×10^{-16}	dynamic 1.38×10^{11}; stale 1.32×10^{-3}
Controlled-output chronology	3	0.0173	2.66×10^{-16}	dynamic 1.70×10^{11}; stale 2.88×10^{-3}
Original fixed-grid eta slope	2	0.49	—	finite-grid, non-asymptotic
Original fixed-grid eta slope	3	0.40	—	finite-grid, non-asymptotic
Small-step audit slope	2	1.99	—	fixed float64 window
Small-step audit slope	3	2.87	—	fixed float64 window

The original 0.0125–0.1 grid probes a finite-step regime and gives slopes $0.49/0.40$. An implementation audit verified that the swept scalar directly multiplies every official fast-parameter update in this controlled one-token map. On a separately fixed float64 small-step grid, fixed-window median slopes are 1.99 for support two and 2.87 for support three; the weakest fitted signal is 341 times its numerical floor. The earlier discrepancy is therefore finite-grid behavior, not a contradiction of the support theorem. This exponent audit applies to the controlled layer map, while the broader positive result is that historical interaction and chronology survive at one predeclared layer of one official checkpoint; ordinary end-to-end inference retains learned token-dependent modulation and composite mini-batches.

R.3 Locked 350M cross-size replication

We independently replicated the same controlled probe on a second official pretrained checkpoint:

```
Test-Time-Training/ttt-mlp-350m-books-2k
revision 0c76ef3883a038c45360737ab8cb0d9e17907c85
  state SHA-256 3c077f537a088edbb2e58ebf618b3810
                1279437d5b3c7cc0eac25cc25a36440a
```

The 24-layer architecture and the same architecture-only rule selected zero-based layer $\lfloor 24/2 \rfloor = 12$. The official JAX implementation from https://github.com/test-time-training/ttt-lm-jax was pinned at commit `6f529b124c7fb5879b33c06926408b15add1d82f`. Seeds 700–704 and all probe definitions were fixed before outcomes; there was no checkpoint, layer, seed, probe, or eta-window search.

Table 6: **Locked 350M pretrained TTT-MLP replication at zero-based layer 12.** Interaction and chronology entries are median Dynamic/Stale floor ratios over five fixed seeds. Every required Dynamic/Stale separation and both slope criteria pass in 5/5 seeds.

Probe	Dynamic / fitted result	Stale / control	Replication
Fast-state interaction, $r = 2$	3.066×10^{11}	4.74×10^{-4}	5/5 seeds
Fast-state interaction, $r = 3$	4.572×10^{10}	3.90×10^{-4}	5/5 seeds
Fast-state chronology, $r = 3$	3.457×10^{10}	3.38×10^{-4}	5/5 seeds
Canonical-response chronology, $r = 3$	1.356×10^{11}	1.30×10^{-3}	5/5 seeds
Controlled-output chronology, $r = 3$	1.429×10^{11}	2.78×10^{-3}	5/5 seeds
Small-step slopes, $r = 2/3$	$1.9884/2.9062$	—	5/5 seeds
Adaptive 2D low/cubic	165.05	random median 52,709.32	319.36×; 32/32

Table 6 gives the complete locked 350M replication summary under the mechanically generalized middle-layer rule.

Dynamic support-two and support-three interactions are far above the numerical floor in every seed, while the corresponding Stale interactions remain near floor. Support-three chronology likewise survives at the fast state, canonical response, and controlled output in 5/5 seeds. The optional mechanism-calibrated, downstream-label-free adaptive subspace uses no downstream examples; its

maximum Stale response is 1.203×10^{-4} floor units. Independent recomputation agreed at stored precision, and all predeclared integrity, finiteness, slope, and subspace checks passed.

We treat this experiment strictly as a cross-size replication; model-size scaling lies outside its scope. Its evidential scope is pretrained mechanistic replication, while spontaneous natural-language use and support-three causal recruitment are separate questions.

S Fixed-Coordinate Scope and Adaptive-Subspace Recovery in Pretrained TTT

The later fixed-channel audit supplies an important coordinate-specific scope boundary. In official default-init TTT-MLP, fast-state low/cubic ratios are 2.186 for a simple swap, 317.10 for w_A, and 148.60 for w_B. The DeepLinear scalar directions therefore fall outside the predefined clean regime under this criterion. Their fixed two-dimensional theory subspace is enriched for cubic signal at default initialization. In the separate fixed-coordinate audit, its low/cubic ratio in the locked pretrained model is 51751, versus 44446 for the median of 64 fixed random two-dimensional subspaces. Signal survival is supported, while specificity of these fixed coordinates is not established.

The mechanism-calibrated, downstream-label-free adaptive extension reuses the 125M checkpoint from Appendix R.2 at zero-based layer 6. A calibration stream estimates Σ_{low} from value degrees zero to two and Σ_3 from degree three in the fixed five-dimensional zero-sum permutation basis. The fixed top-two generalized eigenvectors maximize the calibration chronology/low-order quotient; no downstream labels enter.

On 1,024 disjoint held-out geometries, the adaptive subspace has low/cubic 3494.09, cubic fraction 7.945%, low fraction 0.0442%, enrichment 179.83, and cubic magnitude 7.18×10^9 floor units. It is $13.10\times$ cleaner than the matched-random low/cubic median 45784.03 over 128 fixed random two-dimensional subspaces, ranks above all of them, beats the random median in 32/32 batches, and has stale magnitude 1.21×10^{-4} floor units. This is architecture-derived chronology-signal recovery in learned coordinates, not transfer of the exact DeepLinear coordinates.

T Pretrained TTT: Downstream Scope and Recruitment

T.1 Controlled-language downstream scope

Each episode contains one agent, object, and two rooms. The same actions—go to the object, pick it up if co-located, and return home—appear in all six orders; only go–pick–return succeeds. Semantic actions are independently bijected to abstract event identities, making the successful abstract permutation exactly balanced. A deterministic simulator supplies labels independently of the model. Train/dev/test sizes are $480/120/240$, with a second 240-episode held-out-template test; all label counts are uniform.

The official pretrained model produces six Yes–No score differences from independently cloned trajectories. On held-out entities with familiar wording, raw argmax accuracy is .171; a fixed adaptive two-dimensional linear readout is .254 (95% bootstrap CI $[.204, .308]$), below the random-two-dimensional median .296. The full five-dimensional zero-sum readout reaches .433 (CI $[.371, .492]$). On the held-out paraphrase, the adaptive readout is .0625, whereas the full representation reaches .225. Shuffled-label training gives .150 on the ID test, near six-way chance $1/6$.

These outcomes define a scope boundary. Under familiar wording, task information remains distributed across the full permutation representation rather than being consistently concentrated in the clean mechanism subspace; the held-out paraphrase shows substantial wording sensitivity. The in-distribution full-representation result motivates the hypothesis that behavior may depend on mixtures of historical supports; that hypothesis remains open here.

The following recruitment subsections record the auditable evidence behind the first three parts of Section 5. The model is the official pretrained TTT-MLP 125M at revision `254be4aaff95c3421d7207540b9eb50b468bc0ef`. All recruitment runs start from the same grounded step-650 checkpoint (SHA-256 prefix `9B2E3741`); optimizer state is reinitialized. Training changes model weights. Our reset and transplant interventions are methodologically related

to causal tracing and interchange interventions (Meng et al., 2022; Geiger et al., 2021; 2022); each is applied to a fixed recruited checkpoint.

Training and execution specification. Grounding and D2 recruitment update all model parameters in float32 with target-only causal cross-entropy and reinitialized AdamW ($\beta_1 = .9$, $\beta_2 = .999$, $\epsilon = 10^{-8}$, weight decay .01). Both use exact-length batches of at most 16, no padding, packing, or gradient accumulation, global gradient clipping at 1, and a 50-update linear warmup to 10^{-5} followed by a constant learning rate. Grounding uses seed 2026082802; each D2 seed sees every row of the fixed 8,192-example pool once. The official checkpoint config uses gated shared-Q/K TTT-MLP writes with base inner learning rate .1, 16-token write mini-batches, convolution width 4, and sequential Dynamic state carry. Evaluations and interventions are gradient-free with `use_cache=False`; Stale evaluates every write at the common anchor and Frozen sets both fast-update multipliers to zero.

T.2 Grounding and composition gap

Examples permute three held-out diverse symbols. A depth-one prompt supplies an initial ordering and one permutation; depth two supplies a second permutation but does not expose the intermediate ordering s_1. The grounded R0 checkpoint reaches 98.2% D1 accuracy on the fixed matched test set, while D2 and D3 are 22.2% and 20.8% (six-way chance 16.667%). Supplying the correct s_1 as a new one-step input recovers local competence, localizing the gap to producing and reusing a latent intermediate state rather than applying the operator itself.

T.3 D2-only recruitment replication

Training examples contain (s_0, a_1, a_2) and only the final s_2 target. They never show s_1 and contain no D1, D3, D4, auxiliary, probe, or representation loss. All seeds use the same fixed training pool and held-out symbol pools. The maximum budget is 8,192 D2 examples; no seed outside the predeclared gate is replaced or extended.

For each seed, R^* is the earliest saved checkpoint satisfying D1 $\geq 90\%$, D2 $\geq 80\%$, and nontrivial D2 $\geq 75\%$. Here nontrivial D2 is accuracy on the subset where both composed operations are nonidentity. Results are:

Table 7: Eight-run D2-only recruitment. Seven runs meet the predeclared criterion; one run remains outside it. D3 is never trained. Reset damage and donor-follow change are percentage points at R^*.

Seed	D2 examples at R^*	D1	D2	R^* D3	Final D3	Reset	Donor follow
2026082909	4096	90.4	80.4	30.6	40.0	56.8	67.0
2026082921	4102	93.4	84.4	26.6	45.2	61.2	75.0
2026082922	4096	93.4	85.0	31.0	42.2	61.8	74.6
2026082923	8192	98.4	98.4	43.0	43.0	79.4	96.2
2026082924	4106	95.2	87.4	33.6	39.4	67.0	81.2
2026082925	4096	93.6	85.0	32.6	41.2	62.4	78.4
2026082926	8192	98.4	98.0	37.6	37.6	81.0	98.6
2026082927	—	—	—	—	37.4	—	—
Recruited mean $\pm$ SD	—	94.7 ± 2.9	88.4 ± 7.0	33.6 ± 5.3	41.2 ± 2.5	67.1 ± 9.4	81.6 ± 11.7

Table 7 reports both the auditable R^* D3 values behind the main-text 33.6 ± 5.3 summary and final-checkpoint D3. Seed 2026082927 remains outside the predeclared gate: at its final 8,192-example checkpoint, D1 is 87.6% and D2 is 97.6% (nontrivial D2 97.71%), so it narrowly misses only the predeclared D1-retention threshold and is neither replaced nor extended. Across all eight final checkpoints, D3 is $40.8 \pm 2.7\%$. The reset intervention processes (s_0, a_1) normally, restores the audited task-evolving TTT state to its anchor, and then processes a_2. Every seed has a reset-damage paired-bootstrap interval above zero among the seven recruited models. The corrected transplant uses same-symbol donor/recipient histories with nonidentity first actions and distinct s_1 states; donor s_1 is never printed. Donor-directed score shift is positive with confidence interval above zero in all seven. A separate control uses different nonidentity histories that produce the same s_1; these donor states are more behaviorally interchangeable than different-s_1 donors in all seven.

The fixed linear probe predicts s_1 at 15.8–21.7%, while action identity is decoded at 99.2–100%. Fixed global geometry likewise does not form a clean six-state cluster. These representational diagnostics delimit the claim: the causal conclusion concerns a context/history-dependent intermediate signal, not a globally linear state code.

T.4 Historical-alignment protocol and results

The prespecified extension evaluates historical alignment on every successfully recruited model, with no substitution. Dynamic, Stale, and Frozen use the audited execution modes; Stale evaluates every write from the common initial fast state. The same 500 D1 and 500 D2 examples are paired within each seed, with 10,000 fixed-seed bootstrap resamples.

Table 8: Historical alignment at the earliest recruited checkpoint (percentage points). DiD is $(\text{Dynamic} - \text{Stale})_{D2}$ minus the corresponding D1 gap.

Seed	D1 D–S	D2 D–S	DiD	Support-2 ablation cost
2026082909	10.4	56.8	46.4	61.6
2026082921	15.0	61.8	46.8	70.0
2026082922	16.4	62.8	46.4	69.6
2026082923	9.6	74.6	65.0	80.4
2026082924	19.0	64.6	45.6	75.6
2026082925	17.6	60.4	42.8	70.0
2026082926	8.0	71.2	63.2	78.8
Mean $\pm$ SD	13.7 ± 4.3	64.6 ± 6.2	50.9 ± 9.1	72.3 ± 6.5

Table 8 gives the seed-level Dynamic–Stale, DiD, and support-two ablation results. Support two uses the pre-existing functional definition $I_2 = f_{12} - f_1 - f_2 + f_\emptyset$, with prompts constructed by deleting action instructions while preserving relative order. Removing I_2 from the native candidate-score vector gives the ablated decision. Across the seven models, the mean recruited-model Dynamic ablation cost is 72.3 ± 6.5 points. Relative to the R0 Dynamic ablation cost, its mean increase is 63.89 ± 6.47 points; relative to the Stale ablation cost at the same recruited checkpoint, its mean advantage is 58.23 ± 7.32 points. All seven pass the combined predeclared criterion and all prospective intervals exclude zero.

D2 prompts are longer and usually contain one additional TTT mini-batch, so D1/D2 DiD is not treated as a complete adjustment for chunk count. Same-chunk analyses and the independent reset/transplant controls reduce this concern; the causal claim rests on their convergence rather than DiD alone. The Appendix reports the predeclared protocol and seed-level summaries; corresponding reviewer-facing arrays and per-example margins are included in the anonymous supplementary data. On the frozen matched same-chunk subset (20 D2 prompts, each 48 tokens and hence three 16-token TTT mini-batches), Dynamic–Stale accuracy gaps were $50/45/70/80/50/55/80$ points across seeds (mean 61.4 ± 14.9 points). This is a small conditional diagnostic rather than a fully powered D1/Stale length orthogonalization.

U ControlledTTT Behavioral Use and Acquisition

This appendix records the protocol and seed-level results behind Sections 5 and 6. The purpose of the experiment is behavioral usability, not identification of perturbative coefficients inside the trained network.

U.1 Task calibration and clean interface

The latent state is $(x, y) \in \mathbb{Z}_5^2$. Each episode supplies an initial state, one to four operations drawn from the fixed affine vocabulary, and a query for one final coordinate. Train, development, and test sets contain disjoint operation strings at depths two through four, connecting the task to compositional-generalization tests (Lake & Baroni, 2018; Keysers et al., 2020; Hupkes et al., 2020). A standard GRU trained on the identical interface achieved test accuracies $100.00, 100.00, 98.64, 94.26\%$ at

depths one through four, respectively. Thus the deeper test cases are learnable without memorizing the held-out strings.

The ControlledTTT WRITE call processes the initial state and operations. At READ, those tokens, hidden activations, convolutional halos, and caches are absent; only the explicitly defined fast parameters survive. Dynamic evaluation recomputes each update at the current fast state. Stale evaluation computes all write directions at the episode's initial fast state and then accumulates them. Frozen evaluation disables the writes. These definitions and the checkpoint selection rule were fixed before replication.

Table 9: ControlledTTT architecture and optimization specification.

Component	Fixed specification
Outer backbone	8 blocks, width 256, 4 heads of width 64, gated FFN width 1024; 9,486,381 trainable parameters
Discrete interface	25 state, 6 operation, one padding, and 2 query embeddings; 5 output classes
Fast state	Per block and head, $W_1 \in \mathbb{R}^{64\times 256}$, $W_2 \in \mathbb{R}^{256\times 64}$, and biases; GELU fast MLP; shared Q/K projection with width-1 depthwise Q/K scaling
Inner write	Squared error between layer-normalized fast output and $v-k$; $\eta = .1\max(1+\iota, 0)\sigma(a)/64$, with learned token scalar ι and gate a; one sequential write per valid initial-state or operation token
Outer training	Final-coordinate cross-entropy; batches of 32 with 8 examples per depth; AdamW ($\beta_1 = .9$, $\beta_2 = .999$, $\epsilon = 10^{-8}$, weight decay .01), global clip 1
Schedules	Behavioral runs: 1,200 steps, 50-step warmup to 3×10^{-4}, then cosine decay to 3×10^{-5}. Dense acquisition: 6,000 steps, the same schedule through step 3,000, then constant 3×10^{-5}

U.2 Prospective three-seed replication

The replication used seeds 2026082552–2026082554 without replacement or performance-based extension. Test accuracy for each valid seed was:

Table 10: Dynamic test accuracy (%) for every replication seed.

Seed	Depth 1	Depth 2	Depth 3	Depth 4	Macro
2026082552	100.00	77.38	67.05	48.53	73.24
2026082553	100.00	77.10	67.15	49.03	73.32
2026082554	100.00	65.03	66.88	45.25	69.29
Mean	100.00	73.17	67.03	47.60	71.95
SD	0.00	7.05	0.14	2.05	2.30

Table 10 reports every run in the original prospective three-seed cohort.

Frozen macro accuracy was 20.00%, exactly chance. Stale macro accuracy was 54.57%, while Dynamic reached 71.95%. On the prospectively specified chronology-sensitive subsets, the mean Dynamic-minus-Stale gaps were 25.34, 25.24, and 13.21 percentage points at depths two, three, and four. The ordering Dynamic $>$ Stale $>$ Frozen therefore holds both as a storage control and, on order-sensitive cases, as a recomputation control.

U.3 Functional history-support analysis

For each episode we evaluate the correct-class margin on every retained-history subset, keeping the original relative chronology. Möbius inversion gives

$$I_S = \sum_{R\subseteq S} (-1)^{|S|-|R|} m(R), \qquad m(H) = \sum_{S\subseteq H} I_S.$$

Removing all terms with $|S| = k$ is an exact functional ablation of historical support k for this finite input–output map. Internal-activation interventions and perturbative η-order are separate notions.

Across the three replication seeds, removing support two reduced accuracy by 14.71, 19.01, and 18.75 points at task depths two, three, and four. The support-three contribution at depth three was 7.88 points, and the support-four contribution at depth four was 4.36 points. Each reported contribution had positive sign in all three seeds. Stale evaluation can contain nonlinear higher-support terms,

but those terms did not reproduce the same behavioral benefit. These results support the narrower conclusion used in the main text: task-trained sequential recomputation makes reproducible use of multi-history structure through depth four.

U.4 Fixed eight-run behavioral summary

A later fixed behavioral aggregate contains eight independent ControlledTTT runs with seeds 2026082552–2026082559 under the same task and evaluation definitions. Mean Dynamic accuracy at depths one through four was respectively 99.91%, 70.84%, 66.29%, and 47.03%. Macro accuracy was 71.02% for Dynamic, 52.64% for Stale, and 20.00% for Frozen, with Dynamic $>$ Stale $>$ Frozen in all eight runs. The corresponding support analyses retained positive useful contributions at supports two, three, and four. This behavioral cohort is distinct from the dense acquisition cohort in the next subsection, whose seeds are 2026082561–2026082563 and 2026082569–2026082573. The preceding three-seed table is the locked detailed replication subset for which the full per-depth standard deviations and ablation values are reported here; no unarchived dispersion estimates are inferred for the later aggregate.

U.5 Acquisition: Dense Trajectories and Conditional Timescale Theory

This appendix separates the empirical acquisition result from its analytic explanation. The former concerns end-to-end AdamW trajectories of a controlled TTT learner. The latter is an exact statement for fixed historical features under squared-loss gradient descent.

U.6 Dense-cohort analysis

Eight independently initialized 9.486M-parameter models were trained for 6,000 steps. Eighteen checkpoints per seed were fixed at steps

$$250, 500, \ldots, 3000, 3500, 4000, 4500, 5000, 5500, 6000.$$

At each checkpoint, the decomposition from Section U.3 gives the behavioral contributions $A_2(t)$, $A_3(t)$, and $A_4(t)$. Within each seed we set $N_r(t) = A_r(t)/A_r(6000)$ and define $t_q(r)$ as the first stable crossing of q. Stability requires the immediately next saved checkpoint also to remain above the threshold (with step 6,000 exempt). Positivity times use 10,000 paired bootstrap replicates and the lower endpoint of a 95% percentile interval.

The robust result concerns middle/late recruitment. Seven of eight runs satisfy $t_{50}(2) < t_{50}(3) < t_{50}(4)$, while all eight satisfy the same ordering at t_{75}. Aggregate t_{50} values are $1500, 2000, 2750$ and aggregate t_{75} values are $2000, 2500, 3500$. These aggregate times are computed by averaging the eight seed-level normalized trajectories at each checkpoint and applying the same next-checkpoint crossing rule; they are not medians of the per-seed crossing times. The median ordinary Dynamic-macro halfway time is 1,000. Six runs meet the predeclared strong-nonmonotonicity criterion, so the result concerns stable threshold crossings rather than smooth stagewise trajectories.

Table 11: Stable recruitment times for all eight dense trajectories. "NM" marks that at least one support trajectory has either an adjacent normalized decline larger than 0.25 or at least two total downward recrossings of the predefined $0.25/0.50/0.75$ thresholds.

Seed	$t_{50}(2)$	$t_{50}(3)$	$t_{50}(4)$	$t_{75}(2)$	$t_{75}(3)$	$t_{75}(4)$	NM
2026082561	1500	2000	2750	2000	2500	5000	no
2026082562	1750	2250	3000	2000	3000	3500	yes
2026082563	1500	2000	3000	2000	3000	3500	no
2026082569	1500	2000	2750	2250	2500	3500	yes
2026082570	1250	1750	2250	1750	2500	3500	yes
2026082571	1500	1750	2500	2000	2500	4000	yes
2026082572	1250	2000	3000	1750	2500	4000	yes
2026082573	2000	2000	3000	2250	3500	4000	yes

Table 11 lists all eight dense-trajectory seeds and their predefined stable crossings; these are the acquisition cohort identified at the end of Appendix U.

U.7 Output-level bridge audit

The same eight runs contain archived per-example support decompositions at all 18 checkpoints. Two precommitted output-level diagnostics—correct-margin functional RMS and the leading eigenvalue of centered-logit covariance—satisfy $r = 2 > 3 > 4$ in every seed/checkpoint comparison ($144/144$ for each diagnostic). The hierarchy is widest early and remains strict at step 6,000. These are behavioral-output quantities, not the theorem's frozen feature covariance. A third diagnostic restores parameters, optimizer state, and RNG at checkpoints $250, 1500, 2000, 3000, 6000$ and constructs, but does not apply, the exact next mixed-depth AdamW update $\Delta\theta_t$. For example x, let $m_x(H,t)$ be the correct-class logit minus the mean of the four incorrect logits under retained-history subset H, let $I_S(x,t)$ be its Möbius coefficient, and define the support-r correct-margin contribution $i_r(x,t) = \sum_{|S|=r} I_S(x,t)$. The next-step gain is the signed directional derivative $G_r(t) = \mathrm{mean}_x\, D_\theta i_r(x,t)[\Delta\theta_t]$ on the first 64 archived depth-four examples. The audit definition above is the authoritative reviewer-facing specification. Seed-level diagnostics, complete archived arrays, and the 40 fixed seed/checkpoint records are included in the anonymous supplementary data. The local AdamW diagnostic yields no consistent cross-support ordering, so the present bridge is supported at the output level while optimizer-level ordering remains unresolved.

U.8 Exact frozen-mode dynamics

Mode- and scale-dependent learning dynamics provide the broader context (Saxe et al., 2014; Pezeshki et al., 2022). Let a centered scalar historical feature z_r^η be frozen during outer training, and fit its coefficient by population squared loss using a common gradient-descent step size α. If

$$z_r^\eta = \eta^r g_r + O_{L^2}(\eta^{r+1}), \qquad \mathbb{E}[g_r^2] = c_r > 0,$$

then

$$\lambda_r(\eta) = \mathbb{E}[(z_r^\eta)^2] = c_r \eta^{2r} + O(\eta^{2r+1}).$$

For an aligned noiseless target, the coefficient error obeys $e_{t+1} = (1 - \alpha\lambda_r)e_t$. The exact time required to learn fraction $q \in (0,1)$ of that mode is

$$t_q(r,\eta) = \left\lceil \frac{\log(1-q)}{\log(1-\alpha\lambda_r(\eta))} \right\rceil = \frac{-\log(1-q)}{\alpha c_r} \eta^{-2r}[1 + O(\eta)].$$

The vector/subspace version replaces c_r by the relevant nonzero eigenvalue of the leading covariance. With a nonorthogonal lower-order block, the same result holds after residualization when the Schur complement remains nondegenerate.

U.9 DeepLinear specialization

The exact DeepLinear strength formulas supply the leading constants rather than changing the exponent. With the notation of the DeepLinear appendices, for $r = 2m$ and $r = 2m+1$ the fresh-query mode covariances have leading constants

$$c_{2m} = \frac{\|\beta_0\|^2}{d} \lambda_d^m, \qquad c_{2m+1} = \frac{\mathrm{tr}(H_0^2)}{d^2} \lambda_d^m.$$

Consequently, under the same fixed-feature and common-GD assumptions,

$$t_q(2m,\eta) = \frac{-\log(1-q)d}{\alpha \|\beta_0\|^2 \lambda_d^m} \eta^{-4m}[1 + O(\eta)],$$

$$t_q(2m+1,\eta) = \frac{-\log(1-q)d^2}{\alpha\, \mathrm{tr}(H_0^2) \lambda_d^m} \eta^{-(4m+2)}[1 + O(\eta)].$$

Thus the known key-geometry factor λ_d and fresh-query $1/d$ factors sharpen the constant while historical support fixes the small-η exponent.

U.10 Scope and counterexample

Strict slowing with order follows for any fixed finite set of orders when the leading constants are bounded and nonzero and η is sufficiently small. It is not an unconditional finite-η law. For orthogonal

unit-variance x, y at $\eta = 0.1$, the explicit construction $z_2^\eta = \eta^2 x$ and $z_3^\eta = 100\eta^3 y$ gives covariance scales 10^{-4} and 10^{-2}; with common outer step one and $q = 1/2$, the exact frozen-mode times are 6932 and 69. Thus the higher-support historical mode learns roughly one hundred times faster at this finite η while retaining its formal small-η scaling. A fixed-kernel linearization inherits the theorem only when its kernel, residualization, target projection, and spectral gap remain controlled throughout the stated neighborhood. Those are strong assumptions, not properties proved for the end-to-end AdamW experiment. The analytic result therefore supplies a rigorous mechanism qualitatively aligned with the dense trajectories, not their numerical predictor.

V Sign-Symmetry Identifiability Boundary

State-only diagnostics can fail for structural reasons even when sequential computation is present. The following result identifies a symmetry that makes a sign-odd target unrecoverable from sign-invariant trajectory statistics, preventing a null decoder from being read as evidence that no history-dependent structure exists; Section Z.3 gives a finite-sample check.

V.1 Gradient invariance

If $f_\theta(-k) = -f_\theta(k)$ for every θ, differentiation with respect to θ gives

$$\nabla_\theta f_\theta(-k) = -\nabla_\theta f_\theta(k).$$

The residual also reverses:

$$f_\theta(-k) - (-v) = -(f_\theta(k) - v).$$

Their product is invariant, so the complete squared-loss gradient is unchanged by the paired sign flip.

V.2 Distributional involution

Proposition V.1 (Scoped sign-symmetry identifiability boundary)**.** *Under the paired symmetric episode distribution described below, every measurable function of a sign-invariant trajectory statistic has zero conditional mean for the sign-odd product target. In particular, its Bayes normalized squared error is one.*

Proof of Proposition V.1. Let an episode be $E = ((k_1, v_1), \ldots, (k_T, v_T))$, and let $\tau_i E$ flip one selected pair $i \in S$. By assumption, E and $\tau_i E$ have the same distribution.

All states before i are equal on the two episodes. At update i, gradient invariance makes the new states equal. Every later memory is the same, so deterministic induction yields

$$\theta_T(\tau_i E) = \theta_T(E).$$

The product target is anti-invariant:

$$Y(\tau_i E) = -Y(E).$$

For every bounded measurable test function φ,

$$\begin{aligned}\mathbb{E}[Y\varphi(\theta_T)] &= \mathbb{E}[Y(\tau_i E)\varphi(\theta_T(\tau_i E))] \\ &= -\mathbb{E}[Y\varphi(\theta_T)].\end{aligned}$$

Hence this expectation is zero. The same pairing applies to any deterministic trajectory statistic Z satisfying $Z(\tau_i E) = Z(E)$. Applying the identity to all events measurable with respect to Z yields $\mathbb{E}[Y \mid Z] = 0$ almost surely. Under $0 < \mathbb{E}[Y^2] < \infty$ and squared loss, the Bayes predictor from Z is the conditional mean, so its risk is $\mathbb{E}[Y^2]$; the zero readout attains normalized MSE one.

If initialization is random and independent of the episode, condition on its random seed and apply the same proof. A learned initialization fixed before the current episode is covered identically. No small-η or polynomial assumption is used. □

V.3 What is and is not covered

The proof covers arbitrary fixed/asymmetric initialization, arbitrary finite eta and sequence length, any width, the full vector state, β_T, and every measurable function of a sign-invariant trajectory statistic. The same gradient-history argument can cover momentum or weight decay when the optimizer transition consumes only the invariant gradient sequence and sign-invariant hyperparameters, but the stated proposition remains scoped to the sequential-GD setting.

The proof does not cover:

- a predictor with a bias or other non-odd component;
- a distribution not invariant under the independent paired flip;
- initialization coupled to the current episode's sign orientation;
- a readout that receives oriented keys, raw values, or other non-invariant information;
- a target invariant under the flip;
- vector outputs without a specified output-group action.

An independent query unchanged by the memory flip does not escape: (θ_T, q) remains invariant. A nonlinear state-only readout also cannot escape because the paired states are identical.

W Extension to General Sequential Learners

Meta-learned memories, fast parameterization, self-modifying weights, and recurrent linear attention are related fast-state precedents (Santoro et al., 2016; Munkhdalai & Yu, 2017; Munkhdalai et al., 2019; Irie et al., 2022; Katharopoulos et al., 2020). Distribution-shift test-time adaptation instead updates predictors from unlabeled test data (Sun et al., 2020; Wang et al., 2021). For a scalar differentiable predictor and squared loss,

$$\theta^+ = \theta - \eta(f_\theta(k) - v)\nabla_\theta f_\theta(k) = \theta + \eta D_k(\theta) + \eta v L_k(\theta),$$

where $D_k = -f_\theta(k)\nabla_\theta f_\theta(k)$ and $L_k = \nabla_\theta f_\theta(k)$.

Polynomial learners. If D_k, L_k are polynomial in θ, finite composition gives a finite polynomial in historical values and eta. The support-union induction in Section G applies verbatim, replacing the cubic DeepLinear convolution by the finite polynomial convolution.

Analytic and formal learners. If the fields are analytic near the finite reachable trajectory, analytic composition preserves the filtration locally. Algebraically, the same statement holds in the ring of formal power series without a convergence claim.

Smooth learners. For sufficiently C^r fields and fixed finite horizon, mixed responses involving r distinct historical values are $O(\eta^r)$ under local derivative bounds. A smooth function need not equal its Taylor series; we therefore avoid eta-valuation language without analyticity or a formal-series interpretation.

W.1 Why the simple chain is special

Let $L_i = L_{k_i}$, and evaluate derivatives at θ_0. For the first three consecutive square-free histories, the injection-only leading coefficients are

$$c_1 = L_1(\theta_0), \tag{10}$$
$$c_{12} = JL_2(\theta_0)[c_1], \tag{11}$$
$$c_{123} = JL_3(\theta_0)[c_{12}] + HL_3(\theta_0)[c_1, L_2(\theta_0)]. \tag{12}$$

The Hessian term arises from the cross term in the second-order Taylor expansion of L_3. At higher orders, repeated expansion generates chronological derivative trees. For DeepLinear, L_k is linear, so $JL_k = L_k$ and every higher derivative vanishes; all branching trees collapse to the single chain in Proposition 2.3.

Classical rooted-tree and Volterra expansions provide the background machinery (Magnus, 1954; Butcher, 1972; Murua & Sanz-Serna, 2017; Fliess, 1981); here they are specialized to historical-value support and contrasted with the collapsed DeepLinear chain.

X Experimental Protocols

This section records the float64 theorem-check protocols for the DeepLinear validation experiments; pretrained, behavioral, and acquisition protocols are documented separately in Sections Q–U.

X.1 Original interaction-order study

The locked configuration used key dimension $d = 16$, hidden dimension $h = 32$, bias-free predictors, Xavier-uniform initialization, 100 seeds, and global seed 16180. Learning rates were

$$10^{-4}, 3\times10^{-4}, 10^{-3}, 3\times10^{-3}, 10^{-2}, 3\times10^{-2}.$$

The main key regimes were independent spherical keys and a $\rho = 0.5$ correlated construction. Coefficients were extracted using value-sign/Walsh combinations on square-free probes; the formal manuscript definition remains the continuous multi-index coefficient.

The fitted above-floor Dynamic-DeepLinear eta slopes were 2.0000000000 for support two and 3.0000000119 for support three. Static and stale experimental controls were near numerical floor. The reported digits are the locked full-precision values.

X.2 Additional consistency checks

The all-order factor formula was checked against direct operator recurrence through order eight and dimensions up to 64. Direct trajectory extraction at orders two through four was compared to the leading chain. Additional checks covered permutation geometry, the correlated-key calculation, effective-predictor low orders, and exact paired-sign episodes.

X.3 Prospective factor test

The lock fixed $d = 12, h = 7$, normalized deterministic factor energies, orders five and six, $\eta = 0.02$, 20,000 independent episodes, seed 20260820, and a five-percent relative-error tolerance for second moments. The implementation computed C_r by recurrence, not by the alternating closed form. An independent reference reconstruction reproduces every archived aggregate at stored precision.

X.4 Prospective functional test

The lock fixed $d = 11$, 50,000 episodes, seed 20260821, independent spherical write keys and queries, and a five-percent tolerance. Support four used $\|\beta_0\| = 1$; support five used $h = 12$, $A_0 = [I_{11}; 0]$, $w_0 = e_{12}$, so $\beta_0 = 0$ and $H_0 = 2I$. Five hundred episodes per parity were cross-checked against the independent factor-partition formula.

X.5 Distribution summaries

Table 12: Selected prospective distribution summaries. Energy CV is the coefficient of variation of the squared norm or squared output.

Statistic	factor $r = 5$	factor $r = 6$	predictor $r = 4$	predictor $r = 5$
RMS / energy root	0.08693	0.03432	0.63049	1.24857
Median norm	0.04089	0.01419	0.38986	0.77950
10% quantile	0.00760	0.00233	0.14130	0.27796
90% quantile	0.14259	0.05335	0.99586	1.95519
Energy CV	2.25	2.80	1.88	1.86

The predictor RMS entries in Table 12 are square roots of the empirical mean-square values 0.397522 and 1.558925; query-energy distributions are more skewed.

Y Prospective Prediction Locks and Results

Y.1 Factor-space predictions locked before execution

For $d = 12$, $\|A_0\|_F^2 = \|w_0\|_2^2 = 1$,

$$\mathbb{E}\,\|C_5\|_\Theta^2 = 12^{-2} + 12^{-3} = 0.007523148148148148,$$
$$\mathbb{E}\,\|C_6\|_\Theta^2 = 2/12^3 = 0.0011574074074073.$$

The empirical values were 0.007556304472771784 and 0.001177587820535746, giving relative errors 0.4407% and 1.7436%. Both passed the locked five-percent criterion.

Structured support-four configurations were also locked. Keys (e_1, e_2, e_2, e_3) force only the predicted A block to survive with norm 1.5; keys (e_1, e_1, e_2, e_2) force only w_0 to survive with norm 1.445683229480096. Both had zero recorded entrywise error.

Y.2 Functional predictions locked before execution

At $d = 11$,

$$\lambda_{11} = 76/121.$$

The locked and observed mean squares were:

Order	Quantity	Prediction	Observation	Relative error
4	$\|B_4\|^2$	0.394508572	0.397522012	0.7638%
4	$(q^\top B_4)^2$	0.035864416	0.036614802	2.0923%
5	$\|B_5\|^2$	1.578034287	1.558924766	1.2110%
5	$(q^\top B_5)^2$	0.143457662	0.144168935	0.4958%

The maximum pair-recurrence versus factor-partition errors were 6.32×10^{-16} at even order and 2.48×10^{-15} at odd order.

Y.3 Parity, perturbation, and stale control

At $d = 1, h = 2$, unit write keys/query, $A_0 = (1, 0)^\top$, and $w_0 = (0, 1)^\top$, the locked dynamic coefficients were

$$(B_2, B_3, B_4, B_5) = (0, 8, 0, 32).$$

Perturbing w_0 to $(0.01, 1)^\top$ locked the values

$$(0.04, 8.0004, 0.16, 32.0016).$$

Full simultaneous-GD polynomial extraction matched within 7.11×10^{-15}. Stale coefficients at orders three through five were exactly zero. No seed, initialization, or tolerance was changed after observing results.

Z Additional Controls and Explanations

Z.1 Stale-factor support bound

Each factor in the stale construction of Section 2 is affine in all values. Their product is therefore polynomial of total and historical support at most two. Direct extraction at $i < j$ gives

$$B_{ij}^{\mathrm{stale}} = k_i(\beta_0^\top k_j) + k_j(\beta_0^\top k_i).$$

Every term with historical support at least three is exactly zero. Combining this bound with the generic-survival clause of Theorem 3.1 proves the dynamic/stale separation used in Section 2.

Z.2 WHY THE CORRELATED-KEY CONTROL CHANGED LITTLE

At order two,

$$\|C_2\|_\Theta^2 = \|A_0 k_1\|_2^2 + \|w_0\|_2^2 \, (k_1^\top k_2)^2.$$

Correlation changes only the w block directly. The experiment used

$$k_i = \rho u + \sqrt{1-\rho^2}\, z_i, \qquad z_i \perp u,$$

so

$$\mathbb{E}(k_1^\top k_2)^2 = \rho^4 + \frac{(1-\rho^2)^2}{d-1}.$$

At $\rho = 0.5, d = 16$, this is 0.1, compared with $1/d = 0.0625$. With the recorded Xavier energies

$$\mathbb{E}\,\|A_0\|_F^2 = 21.3333, \qquad \mathbb{E}\,\|w_0\|_2^2 = 1.93939,$$

the parameter-free RMS ratio is

$$\sqrt{\frac{21.3333/16 + 1.93939(0.1)}{21.3333/16 + 1.93939/16}} = 1.02470.$$

The stored experimental ratio was 1.0227. The correlated-key perturbation produces only a small expected RMS change because the correlation-enhanced block is a small fraction of total factor energy.

Z.3 SIGN-SYMMETRY STRESS MATRIX

Paired episodes were evaluated under Xavier, intentionally asymmetric, and arbitrary random-candidate initializations; $d \in \{2, 4, 8, 16, 32\}$; hidden sizes 16 and 32; orders two and three; and $\eta \in \{10^{-3}, 3 \times 10^{-3}, 10^{-2}\}$. Across the locked matrix:

$$\max \text{ paired-state error} = 0, \qquad \max \text{ centered state-target covariance} = 0,$$

and the certified normalized MSE was one. The paired construction makes this an exact finite-sample sanity check of the symmetry proof, not an asymptotic probe result.

Z.4 SCOPE OF MAGNITUDE-BASED INTERPRETATION

The fixed-scaffold experiments show rapidly falling interaction norms and rapidly increasing probe gain from supports two to four. The eta/key laws explain signal scale and required rescaling. Conditioning, nuisance interference, regularized-probe behavior, and noise thresholds require separate evidence and remain distinct from interaction existence and magnitude.