EXACT UNDO: INVERSE-BOUND WRITES MAKE RECURRENT STATES PROVABLY REVISABLE

Anonymous authors
Paper under double-blind review

ABSTRACT

Sequence models can write, but they cannot unwrite. Attention appends keys and values that are never removed; delta-rule networks overwrite by decay. Revision—undo a write exactly, in constant-size state, at unbounded nesting depth—has no architectural primitive. We supply one: *inverse-bound writes*, where a token and its inverse share materialized floating-point parameters at the bit level, so executing the inverse composes the exact inverse of the forward write into the state. Exactness is constructive and parameter-free: one round returns the state to $O(\mathsf{u})$ for every parameter value (Theorem 1); K rounds accumulate at most $K\,O(\mathsf{u})$, while a repeated learned mismatch accumulates at first order (Theorem 3); the construction ports into DeltaNet at a proved price (Theorem 6), and a corollary proves exact cancellation *unreachable* in that family's open parameter domain; measured, its depth-128 undo stacks collapse (0.207) where the orthogonal carrier stays exact. Closure under inverses further makes the state an exact $O(1)$ solver for group word problems, with zero-error $8\times$ length extrapolation. Measured on a group-program benchmark, the taxonomy orders itself as the theory demands: append-only transformers plateau at 0.320; approximate-inverse DeltaProduct shares capability but not cancellation; inverse-closed-but-unbound OS reaches exactness with 5–$10\times$ more samples; inverse-bound OS is exact at the matched schedule, its state exact from initialization, through depth-128 undo stacks. What compute buys is the readout channel—a linear probe decodes depth-128 answers from the unbound model's contracted signal at 1.000 while its first-layer operator is off by 0.8, an error the trained readout's margin tolerates; what the binding guarantees, in the vocabulary-bound layer, is the whole operator, every direction, any probe. A dose–response boundary chart closes the paper: each inverse-bound layer monotonically degrades byte-level language modeling—composition, not compression. The scope is the strictly nested (LIFO) core of revision.

1 INTRODUCTION

Roll back a transaction. Undo an edit. Backtrack a search. Close a bracket. Revision is a primitive of every sequential computing system, and it is an algebraic one: *to revise is to compose a write with its inverse*. Its native form is nested and last-in-first-out, and modern sequence architectures have no primitive for it. Attention is an append-only log: every key–value pair is kept forever, and revision is simulated by re-reading the history and softly overwriting it. Delta-rule fast-weight networks overwrite by decay: erasure is not inversion—nothing composes back to the identity. The write/undo symmetry is absent from these operator families, while the demand for it is documented: stale-context overrides are a live failure mode (Xu et al., 2024); KV-cache editing is an inference-time patch on an append-only substrate (Li, 2026; Moschella et al., 2026).

We restore the symmetry at the vocabulary level. In an *operator stream* (OS), the recurrent state is itself a composed operator: token a multiplies the state by a learned rotation $R(a)$ and adds a translation $v(a)$, so the state after a sequence is the ordered product of the writes. We pair every token a with an inverse token a^- whose write is *by construction* the exact inverse of a's write: the two tokens share the same materialized rotation entries at the bit level ($\cos$ identical, $\sin$ negated), and the inverse translation is computed as $-R^\top v$. Executing a then a^- composes $R^{-1}R = I$ into the state. *Undo is a write*: it costs one step, lives in the same operator family, and its exactness does not depend on what the network has learned. We call this *inverse binding*. It lifts the self-inverse

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.

binding of vector-symbolic architectures (Kleyko et al., 2022) from feature vectors to the transition operators of a recurrent network.

The guarantees are constructive—zero parameters, zero training cost, holding for every parameter value—not training observations. Theorem 1 bounds the undo error by $c_1 \mathsf{u} \|P\|_2$ for *all* parameter values and states; the dual statement is an exact identity: without binding, the undo error *is* the learned inverse mismatch $(\widehat{M} - M(a)^{-1})M(a)$, whatever the learning has made it. Theorem 3 separates the regimes over long horizons: K do/undo rounds cost at most $K\,O(\mathsf{u})$ parameter-free, while a repeated learned mismatch accumulates systematically at first order; measured error comes in far under the bound—sublinear and saturating (Figure 5)—because $\mathrm{fl}(\cos)^2 + \mathrm{fl}(\sin)^2$ rounds back to 1 exactly (Appendix C).

Scope of the guarantee. Executing a^- composes the inverse of the *most recent unmatched* write: exact revision here is LIFO—well-bracketed nesting, at unbounded depth. Undoing an earlier, buried write requires conjugating by the suffix product, which constant-size state does not carry; deciding *what* to undo is addressing machinery on top of the recurrence. We take this as a division of labor: inverse binding supplies the exact algebraic primitive—the inner loop of rollback, nested transactions, bracketed edits—while selective, out-of-order revision is an addressing problem (one that attention solves by re-reading its log at every step, the opposite end of the design axis). This paper delivers and prices the primitive; the theorems cover its strictly nested core.

Nor is the mechanism married to our carrier. DeltaNet's write $A = I - \beta kk^\top$ is invertible off the singular face $\beta \|k\|_2^2 = 1$, and Sherman–Morrison yields an undo write *in the same functional family*, $\beta' = -\beta/(1-\beta)$ (Theorem 6; verified to machine precision, gated and multi-head variants unit-tested)—at a provable price: the undo step must counter-decay, $\|A'\|_2 = 1/|1-\beta| > 1$, where the orthogonal family undoes isometrically.

Closure under inverses buys capability, not just hygiene: a recurrent network whose transition operators are closed under inversion is an exact $O(1)$-state solver for the word problems of every group its operators realize—free reduction over generators is the word problem of the free group, in Dehn's sense (Dehn, 1912). This yields the classification axis Section 4 measures on: *append-only*, *approximate-inverse*, *inverse-closed*.

Our contributions:

- **Inverse-bound writes** (Section 2): to our knowledge the first token-level mechanism making recurrent state exactly revisable *by construction*—bit-reused inverse writes, distinct from global architectural invertibility used for memory and gradients (Gomez et al., 2017; Behrmann et al., 2019; MacKay et al., 2018) and from learned stacks (Section 5).
- **Constructive, parameter-free guarantees** (Theorems 1, 3): undo error $c_1 \mathsf{u} \|P\|_2$ for *all* parameter values and states; K rounds at $K\,O(\mathsf{u})$ vs. first-order systematic accumulation of a repeated learned mismatch; the unbound dual is an exact identity.
- **A measured capability taxonomy** (Section 4): append-only transformers, approximate-inverse DeltaProduct—whose exact cancellation is unreachable in-domain (Corollary 7)—inverse-closed-but-unbound OS, which reaches exactness given 5–10$\times$ compute, and inverse-bound OS, exact at the $1\times$ schedule with the state exact from initialization.
- **Portability with a price tag** (Theorem 6): exact undo inside DeltaNet ($\beta' = -\beta/(1-\beta)$, Sherman–Morrison), three quantified costs, nested undo compounding exponentially in depth (measured 1.7×10^{11} vs. 3.0×10^{-6} at depth 64); $21/21$ unit tests, wrong variants separated by up to 5.5×10^{13}.
- **A charted boundary**: a dose–response study over $0/2/4$ inverse-bound layers in a byte-level LM degrades monotonically in every seed—composition, not compression.

2 The Mechanism: Operator Streams with Inverse-Bound Writes

2.1 Operator streams

The state of an OS layer is an affine operator (P_t, c_t) with $P_t \in \mathbb{R}^{r\times r}$, $c_t \in \mathbb{R}^r$. Token x_t emits a single-plane Givens rotation $R_t = R(\theta_t)$ (materialized as I plus a rank-2 modification on a learned

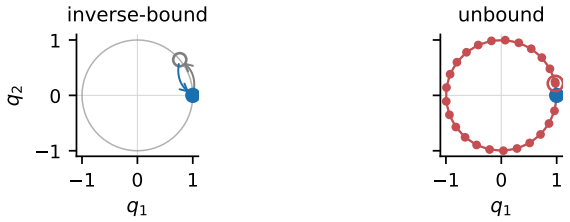


Figure 1: The mechanism in one picture (fp32 simulation; $\theta = 0.7$, mismatch $\delta = 0.25$ rad for visibility; probe $q_0 = (1, 0)$, K=26 do/undo rounds). *Left, bound:* the undo reuses the forward write's stored entries; every round returns to q_0 exactly (drift 0; Theorem 1). *Right, unbound:* the inverse symbol carries a mismatch δ; the state walks off by the learned inverse mismatch of Proposition 2.

plane), a write vector v_t, and a read probe q_t; the recurrence is

$$P_t = R_t P_{t-1}, \qquad c_t = R_t c_{t-1} + v_t, \qquad y_t = P_t q_t + c_t. \tag{1}$$

In homogeneous coordinates $M = \left(\begin{smallmatrix} R & v \\ 0 & 1 \end{smallmatrix}\right)$, a sequence of tokens composes to the ordered product of its writes, and for orthogonal R the inverse is closed-form: $M^{-1} = \left(\begin{smallmatrix} R^\top & -R^\top v \\ 0 & 1 \end{smallmatrix}\right)$. The state lives in the affine group; composition *is* matrix multiplication.

Because composition is associative, the recurrence parallelizes: prefix products $A_t = R_t \cdots R_1$ by log-doubling, orthogonality turning the translation scan into $c_t = P_t \sum_{s \le t} P_s^\top v_s$. The parallel form matches the incremental form to $\le 7.6 \times 10^{-17}$ in fp64 (causal prefixes bitwise-identical) and speeds up training 12–13× (Appendix B). Inference state is constant-size (17 KiB per sequence), independent of context length.

2.2 Inverse binding

Let the vocabulary carry a fixed-point-free involution $a \mapsto a^-$ (each generator gets a designated inverse symbol). *Inverse binding* is the implementation contract under which the write of a^- is the exact inverse of the write of a:

Assumption A (inverse-binding contract). The token embedding of a^- reuses $\mathrm{rep}(a)$, and the write of a^- is constructed at runtime from that of a by (i) $R(a^-)$ has matrix entries equal to those of $R(a)$ up to sign (cos entry identical, sin entry negated)—a bit-level copy; (ii) $v(a^-) = \mathrm{fl}(-R(a)^\top v(a))$.

The cost is zero parameters and one fused sign-and-transpose at emission time. The crucial property is (i): forward and inverse tokens use the *same materialized floating-point entries*. Large θ can make $\mathrm{fl}(\cos\theta)$ far from $\cos\theta$, and the undo does not care: it cancels *the write that was actually executed*, not the ideal write (Figure 1). All our bounds are *as-executed* statements (Section 3). Assumption A governs one layer's write pair, and holds by construction wherever that layer's input at the undo position is the same token representation as at the do position—the first block of our binding models, which read the shared embedding without positional encoding. Deeper blocks are contextualized and outside the theorems; the deep-nesting results of Section 4.4 are task-level.

Assumption H1 (uniform materialization bounds). The stored entries of a write arise from a single angle: there exists $\bar\theta$ with $\hat c = \mathrm{fl}(\cos\bar\theta)$ and $\hat s = \mathrm{fl}(\sin\bar\theta)$ (the implementation evaluates one angle per write and stores both libm results). In particular $|\hat c|, |\hat s| \le 1$, hence $\left\|\widehat{R}\right\|_2 \le \sqrt{2}$ for all $\bar\theta$. Constants in all theorems depend on induced operator norms and computation-graph depth, not on the magnitude of $\bar\theta$. The single-angle provenance is what the orthogonality analysis uses (Appendix A.2, Lemma A.2); the value-range clause alone would not suffice. *Plane separation:* the two pre-orthogonalization directions d_1, d_2 of a write satisfy $|\langle d_1, d_2\rangle| \le (1 - \Delta)\, \|d_1\|_2 \, \|d_2\|_2$ for a fixed $\Delta \in (0, 1]$; the basis constants carry $1/\Delta$, and the anchored-axis construction keeps Δ bounded away from zero by design (Appendix A.2).

Table 1: Transition families and their in-family inverses. "Revisable" is Definition 4.

Family	Write	In-family inverse	Revisable
Attention KV cache	append (k, v)	none (free monoid)	No
Scalar decay (GLA/LRU)	$S \mapsto aS + B$	$(1/a, -B/a)$, $a \neq 0$	Algebraic
DeltaNet	$(I - \beta kk^\top)S + \beta kv^\top$	Theorem 6	Off singular face, at a price
Operator stream (ours)	$(P, c) \mapsto (RP,\ Rc{+}v)$	$R^\top$, $-R^\top v$	**Yes, by binding**

2.3 Writes across families

Inverse binding turns a family-level algebraic property into a vocabulary-level mechanism. Table 1 locates the studied families.

3 Theory

Floating-point model: standard round-to-nearest with unit roundoff $\mathsf{u} = 2^{-24}$ (fp32), $\gamma_n := n\mathsf{u}/(1-n\mathsf{u})$, and $|\mathrm{fl}(AB) - AB| \leq \gamma_n |A||B|$ elementwise (Higham, 2002). Norms are spectral.

3.1 Constructive exact undo

Theorem 1 (Constructive undo)**.** *Let Assumptions A and H1 hold. For arbitrary network parameter values, arbitrary state (P, c) and token a, let $(\widetilde{P}, \widetilde{c})$ be the floating-point state after executing the adjacent pair (a, a^-). Then there exist constants c_1, c_2 depending only on r, the materialization depth k_R (of the stored-entry arithmetic graph), the induced operator bound, and the plane separation Δ of H1, such that*

$$\left\|\widetilde{P} - P\right\|_2 \leq c_1 \mathsf{u}\ \|P\|_2\,, \qquad c_1 = O\big(r\,k_R + r/\Delta\big), \tag{2}$$

$$\|\widetilde{c} - c\|_2 \leq c_2 \mathsf{u}\,(\|c\|_2 + \|v(a)\|_2), \qquad c_2 = O\big(r\,k_R + r/\Delta\big). \tag{3}$$

The constants are independent of the numerical values of the parameters.

OS-reachable states satisfy $\|P\|_2 \leq (1 + c_0\mathsf{u})^t$, so for any realistic $t\mathsf{u} \ll 1$ the absolute error is $O(\mathsf{u})$; the relative form is the natural statement of the proof (Appendix A).

Proposition 2 (Unbound dual: exact identity)**.** *Without binding, let a^- be independently parameterized as an arbitrary operator $\widehat{M}$. Then, exactly,*

$$\widehat{M}\,M(a) - I_{r+1} \;=\; \big(\widehat{M} - M(a)^{-1}\big)\,M(a)\,:$$

the undo error is a linear function of the learned inverse mismatch—precision does not enter the mismatch term (execution adds the baseline $O(\mathsf{u})$ of Appendix A)—and the mismatch vanishes if and only if the learned inverse is exact.

3.2 Drift separation across rounds

Theorem 3 (Drift separation)**.**

(a) Exact identity, block-paired rounds. *For K adjacent do/undo pairs, exactly $\prod_{k=1}^{K} \widehat{M}_k M_k = \prod_{k=1}^{K}(I + F_k)$ with $F_k := (\widehat{M}_k - M_k^{-1})M_k$, for every realization of the mismatches; under binding, $\|F_k\|_2 = O(\mathsf{u})$ (Theorem 1).* ***(b) Repeated mismatch accumulates systematically.*** *If the same pair recurs under the same parameters and context ($F_1 = \cdots = F_K = F$, $\delta = \|F\|_2$—deterministic networks make this the realistic recurring case: calls/returns, open/close, transaction/rollback), then for $K\delta \leq \frac{1}{2}$, $\big\|(I + F)^K - I\big\|_2 \geq K\delta(1 - K\delta)$: cancellation cannot occur at first order, the first-order term being KF.* ***(c) Constructive bound, bound side.*** *Under binding, $\|P_K - P_0\|_2 \leq [(1 + c_1\mathsf{u})^K - 1]\ \|P_0\|_2 \leq K c_1 \mathsf{u}\, e^{K c_1 \mathsf{u}}\ \|P_0\|_2$—machine precision, K, and data scale only, not the accuracy of any learned inverse. (Proofs in Appendix A.4; Figure 5 and Appendix C give the measured curves.)*

3.3 WORD PROBLEMS: CAPABILITY AND COST

Definition 4 (Revisable family)**.** A transition family $\mathcal{M}$ is *revisable* if the monoid it generates under composition is a group—which holds as soon as the inverse of every one-step write is again a one-step write of the family. (The family need not be composition-closed: products of single-plane writes are not single-plane.) Inverse binding upgrades revisability from a family property to a vocabulary-level architectural mechanism.

Proposition 5 (Word-problem solving under closure)**.**

(a) Every finite group G has a faithful orthogonal representation $\rho : G \hookrightarrow O(r)$ ($r \leq |G|$; S_n: $r = n$; A_5: $r = 3$). Taking vocabulary $V = G$ with token operators $R(g) = \rho(g)$, $v = 0$, and inverse binding, the recurrence $x_{t+1} = \rho(g_t)x_t$ implements word multiplication exactly in exact arithmetic, with end-to-end error $O(T\mathsf{u})$ under stable orthogonal products. For $A_5 \subset SO(3)$ this is constructive: all 60 icosahedral rotations are single-plane writes (u, w, θ), verified to 1.3×10^{-15} (Appendix B). The proposition is a statement about the operator family; Section 4 measures what training reaches, and closure is sufficient, not necessary—an unbound family that learns *exact inverses also solves it, at the contingency price Proposition 2 quantifies.*

(b) ***Cost and complexity backdrop*** *(cited, not proven): single-plane writes cost $O(Tr^2)$ on the materialized state ($O(Tr)$ for axis-aligned planes; our same-plane control), against $\Theta(T^2)$ for an append-only store that re-reads its log per query. Word problems of nonsolvable groups (incl. S_5) are NC^1-complete (Barrington, 1989), parity is not in AC^0 (Furst et al., 1984; Håstad, 1986), and log-precision transformers sit in uniform TC^0 (Merrill & Sabharwal, 2023)—the difficulty backdrop of the S_5 testbed, matching the framing of Siems et al. (2025).*

3.4 PORTABILITY: EXACT UNDO INSIDE DELTANET

Theorem 6 (Portability)**.**
(a) Exact undo write. *DeltaNet executes $S_t = AS_{t-1} + B$ with $A = I - \beta kk^\top$, $B = \beta kv^\top$, $\|k\|_2 = 1$ (as implemented). For $\beta \neq 1$, the map $S_{t-1} \mapsto S_t$ is undone exactly by one step of the* same functional family*:*

$$k' = k, \qquad v' = v, \qquad \beta' = -\frac{\beta}{1-\beta}, \tag{4}$$

and $B' = -A^{-1}B$ holds automatically (Sherman–Morrison: $A^{-1} = I + \frac{\beta}{1-\beta}kk^\top$); without $\|k\|_2 = 1$: $\beta' = -\beta/(1 - \beta\|k\|_2^2)$. In floating point the undo error is $O(\mathsf{u}/|1-\beta|)$. ***(b) Singular face.*** *Invertibility fails only on $\beta\|k\|_2^2 = 1$ (zero eigenvalue along k); the family is invertible on an open dense set.* ***(c) Price of the port, in three parts.*** *(i)* Domain*: for every $\beta \neq 2$, β' lies outside the sigmoid-based forward parameterization $\beta \in (0, 2)$—porting requires a structural write bypassing the parameterization, not a clamp fix. (ii)* Expansion*: the undo step has eigenvalue $1/(1-\beta)$ along k, of modulus > 2 iff $|1-\beta| < \frac{1}{2}$, diverging as $\beta \to 1$. (iii)* Stability tension*: $\|A'\|_2 = 1/|1-\beta| > 1$ always violates the constructive $\|A\|_2 \leq 1$ stability of the family—an undo must counter-decay—whereas the orthogonal family undoes isometrically, $\left\|R^{-1}\right\|_2 = 1$, at no stability cost.* ***(d) Gated variant.*** *The gated recurrence ($\alpha = \exp(g) > 0$ multiplying only the old state, as in the reference kernel) admits a* unique *same-form undo—$\tilde{v} = v/\alpha$, $\alpha' = 1/\alpha$—whose full statement, uniqueness proof, kernel convention, and measurements are in Appendix D.* ***(e) Multi-head.*** *Undo of sequential composition must be executed in reverse head/step order (time reversal of a product); $n_h \geq 2$ is an implementation detail, not a condition of invertibility.*

Depth compounds the port's price. Theorem 6(c) is per-step, and nesting compounds it: a bracketed DeltaNet undo of depth d passes, in exact arithmetic, through the historical states, but each undo step's $O(\mathsf{u})$ error is amplified by the remaining undo norms $\kappa = \|A'\|_2$, giving $O(d\,\kappa^d\mathsf{u})$—exponential in depth, where the orthogonal carrier amplifies by 1 and pays $O(d\mathsf{u})$ (measured; Remark A.9).

Corollary 7 (Exact cancellation is unreachable in the open domain)**.** *For any two nonempty finite words W, W' of in-domain DeltaNet/DeltaProduct writes ($\beta_i \in (0, 2)$, $\|k_i\|_2 = 1$), $W'W = I$ is*

impossible. Each write is non-expansive with $\left\|I - \beta kk^\top\right\|_2 = 1$*; $W'W = I$ with* $\|W\|_2, \|W'\|_2 \leq 1$ *forces W to be an isometry, and an invertible product of such writes is an isometry only if every factor has* $\beta_i \in \{0, 2\}$*—both endpoints excluded by the open parameterization* $\beta = 2\sigma(z) \in (0, 2)$ *(Theorem 6(c)(i) is the single-write face). Exact cancellation is therefore not a measure-zero target but an* unreachable *one, approachable only in the limit* $\beta \to 2$*; the gated variant contracts by* $\alpha \leq 1$ *a fortiori.*

As-executed caveat. In fp32, $2\sigma(z)$ rounds to exactly 2 for $z \gtrsim 16.6$, which would realize the excluded endpoint (a Householder reflection is an involution); measured, the archived DeltaProduct checkpoints (three seeds, 20k steps; `diag_beta_check`) execute $\max \beta = 1.9970$ with $\max |z| = 6.5$, and the undo-DeltaNet runs hold $\min |1 - \beta| \in [0.40, 0.53]$ (Appendix B): every executed write stays interior.

4 Experiments

All numbers are script-generated and archived; every figure and table cites its result file. Hardware: RTX 3060 Laptop (fp32 training, fp64 CPU verification); full protocols in Appendix B.

4.1 The taxonomy, measured

Fixed-length compositional programs over a group vocabulary (noncommuting generators; the label is the ordered product; the generated group is A_{12}, nonsolvable—Appendix B). Table 2 is the paper in one table: rows are architectures ordered by how close their write family comes to closure under *exact* inverses, columns are the two schedules, and the theory's prediction—capability follows representability of the inverse—is the row order.

Append-only. Transformers plateau at 0.320 ± 0.011 (matched-position 0.317 ± 0.008) with no grok transition within 20,000 steps at matched depth, consistent with the fixed-depth TC^0 ceiling (Merrill & Sabharwal, 2023); we do not claim the gap is uncloseable by depth alone.

Approximate-inverse. DeltaProduct (n_h=2 negative-eigenvalue Householder heads; 65,744 parameters, fewest of all contestants) shares the plain capability at $10\times$: 0.993 ± 0.007. What it cannot reach is cancellation exactness. At L=8 its cancel and pure-chain scores (0.986 ± 0.000, 0.982 ± 0.007) sit within one standard deviation of its own plain score—shallow stacks carry no differential—while depth-128 stacks collapse to 0.207 ± 0.050 even at the $10\times$ schedule: the depth-compounded face of Corollary 7, under which the exact inverse is unreachable in the open domain, approachable only as $\beta \to 2$ (measured β presses to 1.9970, Section 3.4; each nested undo compounds the shortfall, Remark A.9).

Inverse-closed, learned. The unbound OS carries the exact inverse in-domain (the transpose of a rotation is a rotation) but must learn it: 0.885 ± 0.082 at the $1\times$ schedule, 0.999 ± 0.001 by the 10k checkpoint, 1.000 at 20k. The carrier makes exactness *learnable*, at a 5–$10\times$ sample premium over the bound twin.

Inverse-bound. The bound OS is exact in every cell at the $1\times$ schedule, zero seed variance, and its state is exact from initialization (Section 4.3): binding makes exactness *free*. The sharpest evidence that this is construction rather than learning: one seed's plain-program accuracy collapses at lengths 17 and 257 (0.816, 0.098) while the same weights resolve depth-128 undos exactly (Section 4.4).

Two controls dissect the OS channel (Appendix B): *same-plane* lands on the order-blind predictor, *zero-angle* on the query-conditional marginal—the ordered, noncommuting channel is load-bearing. A plain DeltaNet (n_h=1; Section 4.5) sits in the middle tier at 0.300; each token writes one plane, so a two-plane operation occupies two tokens (Appendix B).

4.2 Zero-shot length extrapolation: eight times zero-error

Trained only at $L = 16$ (post-grok, step 7250), evaluated zero-shot at longer horizons (Figure 4, right, Appendix B): accuracy stays *exactly* 1.000 through $8\times$ the training length, with decay onset near $T \approx 256$ consistent with linear accumulation $T \cdot \left\|\widehat{R} - R\right\|_2$ of a learned-but-inexact single-step operator.

Table 2: The taxonomy, measured: exact accuracy, mean±std over 3 seeds, identical data streams and optimization ($1\times$ = 2,000-step matched schedule; $10\times$ = 20,000 steps; L=8 programs except the depth-d undo-stack columns; `diag_controls`, `diag_sched10x`, `diag_deepnest`, `diag_deepnest_20k`). "state resid." = end-of-training undo-pair deviation of the family's state ($\|P_2 - I\|_2$ for OS; end-state norm relative to a single forward write for DeltaProduct; deep-nest mixed protocol). Transformer stack cells not run: plain programs already plateau at 0.320. Pure-chain columns: $0.982\pm.007$ (DeltaProduct), $1.000/1.000$ (unbound/bound). The bound OS row uses the 86,084-parameter configuration (Appendix B); at the equal-parameter 87,620 configuration it is likewise exact on all three families (Table 6).

Model (tier)	plain $1\times$	plain $10\times$	cancel $10\times$	d=128 $1\times$	d=128 $10\times$	state resid. $10\times$
Transformer (append-only)	0.201	$0.320\pm.011$	$0.441\pm.002$	–	–	–
DeltaProduct (approx. inv.)	0.390	$0.993\pm.007$	$0.986\pm.000$	$0.253\pm.074$	$0.207\pm.050$	0.60
OS unbound (inv.-closed)	$.885\pm.082$	1.000	$.999\pm.002$	$.615\pm.302$	1.000	$.094\pm.011$
OS bound (inverse-bound)	**1.000**	1.000	1.000	**1.000**	1.000	1.3×10^{-7}

4.3 The binding differential: what compute buys, and what it cannot

Single-variable ablation at *strictly equal parameter count* (87,620 both; a shared 60-row embedding table, of which the bound variant indexes 36 rows): only inverse semantics can explain the difference. At the matched schedule the bound model is exact on all three program families and the unbound model is not ($1.000/1.000/1.000$ vs. $0.885/0.927/0.973$ main/cancel-heavy/pure chains, Table 6); at $10\times$ the unbound model closes every task cell (Table 2). The differential that survives compute is the state: on (a, a^-) pairs the bound residual is flat at 1.2×10^{-7} from initialization on, while the unbound residual peaks at 0.23 as the operation table differentiates and then decays slowly (post-peak tail fit, steps $\geq$2.5k: factor 0.87 per 5k steps; 0.20 at the $1\times$ endpoint, 7.4×10^{-2} at 20k); the decay decelerates, so a distance of $\sim 5 \times 10^5$ steps to the bound's level is a lower bound—at least $23\times$ the $10\times$ schedule (Figure 3; program-task protocol; deep-nest protocol: 0.094 ± 0.011, same regime, Appendix B).

Where the unbound error lives. The task reads the state through a low-rank contraction: position t's signal is $y_t = P_t h_t$ with one direction h_t per head. On the unbound model's first-layer signal y_0, logistic regression decodes the answer at 1.000 held-out accuracy, transferring zero-shot to depth 128 at 1.000 (`diag_localize`): the contracted signal carries the answer. The raw state P_0, taken before the query token, is not linearly decodable for either variant (chance, 0.06–0.08 against 12 classes), as the bilinear readout requires—the answer is $\pi(j)$ for a query object j that P_0 does not contain; even the bound model, whose P_0 is machine-exact, sits at chance, so this probe measures readout geometry, not state quality. Nor is the unbound error hidden in unread directions: at depth 128 its first-layer state is off by 0.80 (max-norm, mean over seeds; growth slope 0.44), and the readout-visible component $\left\|(P_{\text{end}} - P_{\text{core}})h\right\|_{\max}$ is 0.46—over half the error lands in the contraction itself (bound layer: 1.0×10^{-6}), and what keeps accuracy at 1.000 is the margin of the trained readout, not invisibility. In every seed the error is asymmetric across heads—one head 2–6$\times$ cleaner, which head varies by seed. The bound first layer is exact in *every* direction at every depth (1.6×10^{-6} at 128, slope 0.48), by construction, for any probe the network may someday be asked to run. And the bound model itself shows the task never needed operator exactness in contextualized layers: its second layer's state drifts to 0.72 at depth 128 with task accuracy unaffected. *Compute buys a channel the margin can live with; the binding guarantees the operator—in the vocabulary-bound layer, every direction, any probe.*

4.4 Deep-nested undo: the state axis, separated

Well-bracketed programs core, $a_1 \cdots a_d$, $a_d^- \cdots a_1^-$ (a LIFO undo stack of depth d; true value always the core's) make revision *necessary*. Trained on length-8 programs with shallow cancellations only, evaluated at depths up to 128: the bound model is exact at *every* depth and *every* seed at the $1\times$ schedule ($21/21$ cells at 1.000, zero variance)—in the vocabulary-bound layer the stack cancels by construction, so undo depth never enters that state (Theorem 1); the model-level cells are task-level evidence (Section 2.2). The unbound model degrades at $1\times$ (0.615 ± 0.31 at d=128) and closes

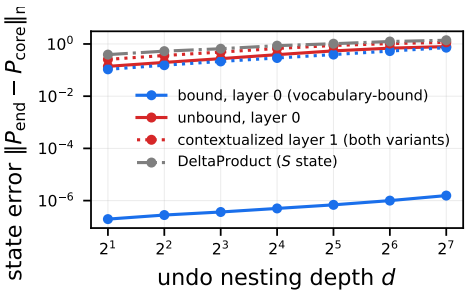


Figure 2: State error vs. undo nesting depth, 20,000-step models (mean over 3 seeds $\times$ 128 programs; `diag_localize`). Bound layer 0 (the vocabulary-bound layer): machine precision at every depth, sublinear growth (1.6×10^{-6} at d=128, slope 0.48). Unbound layer 0: grows to 0.80 (slope 0.44). Contextualized layer 1 (both variants) and DeltaProduct: $O(1)$ or worse. Task accuracy is 1.000 at every depth for both OS variants—the unbound error lands in the readout contraction itself (0.46 of the 0.80 at d=128) and is absorbed by the trained margin (Section 4.3); operator exactness is the binding's guarantee, not a task requirement.

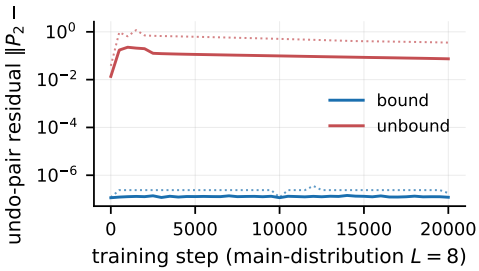


Figure 3: Undo-pair residual $\|P_2 - I\|$ during training, 20,000 steps (48 pairs, all generators; `diag_residual_20k`, program-task protocol). Bound: machine precision at every step, from initialization. Unbound: the residual *is* the learned mismatch—it peaks at 0.23 as the operation table differentiates, then decays slowly at a decelerating rate (post-peak tail fit, steps $\geq$2.5k: factor 0.87 per 5k steps; 0.20 at the 1$\times$ endpoint, 7.4×10^{-2} at 20k), six orders above bound throughout; the bound's 1.2×10^{-7} lies at least $\sim 5 \times 10^5$ steps away, while task accuracy is already 1.000.

the gap at 10$\times$ (21/21 at 1.000) while its end-of-training residual stays at 0.094 ± 0.011 (deep-nest protocol): the task-level gap was compute; the state-level gap is not. DeltaProduct's stacks do not close at either schedule—0.996 at depth 2 decays to 0.207 ± 0.050 at depth 128 even at 10$\times$—the task-level face of the amplification Remark A.9 measures at the mechanism level. Figure 2 separates the axes directly at the 10$\times$ schedule: task accuracy 1.000 at every OS depth cell, against state errors—bound layer 0 flat at machine precision, unbound layer 0 at 0.8, contextualized layers $O(1)$, DeltaProduct highest. The stack is learnable on an inverse-closed carrier; operator exactness is what binding adds—and what no amount of task loss moves.

4.5 Portability: unit tests and a trained undo-DeltaNet

Theorem 6 is verified by construction-level unit tests against the reference recurrence (21/21 green, `test_port`): each acceptance criterion is paired with a deliberately wrong variant that must fail—separated from the correct formula by factors up to 5.5×10^{13}—so the test detects a wrong formula, not just a small number (Table 4, Appendix A.6). The injection also survives the scan engine and training: an adjacent do/undo pair returns the state to zero at 2.4×10^{-7} (`b6`), and at near-equal scale (59,984 vs. 61,520 parameters, the injected model the smaller) it beats its plain sibling on every undo-semantics probe—cancel-heavy programs (1.000 vs 0.726 at L=2; 0.932 vs 0.709 at L=4) and pure chains at L=8 (0.850 vs 0.509, 3 seeds; `diag_undo_dn`). The pure-chain cell is exact at the state level but not at the task level: DeltaNet's output is a learned readout of the state, whereas the OS probe $y = Pq + c$ *computes* a cancelled program's answer in the forward pass—the 0.150 gap is readout learning, not undo drift. Plain programs favor the plain variant (0.300 vs 0.219 at L=8, a 27% relative gap, both pre-grok): the override couples β to the undo requirement—the forward-side face of the price Theorem 6(c) quantifies.

4.6 A charted boundary: composition vs. compression

Where does exactness pay and where does it not? Byte-level language modeling (Python stdlib corpus, 2.0M predicted bytes, identical data stream and optimization): the inverse-bound model trails a same-width transformer (2.8724 vs. 2.3487 val BPB; 860,432 vs. 957,184 parameters), and a dose–response hybrid sweep (0/2/4 OS layers replacing attention layers, all else identical) de-

grades monotonically in every seed—individually, not only on average ($2.356 \pm 0.010/2.729 \pm 0.057/2.840 \pm 0.023$ val BPB over 3 seeds; $\sim$1M parameters, one epoch over 2.0M bytes); the byte vocabulary carries no undo involution, so the binding never triggers and the measured price is the carrier's—the compositional-state parameterization itself. Incremental decode at $B{=}1$ runs at 281 vs. 779 tok/s: the constant-state advantage does not reach wall-clock at this scale (kernel-launch bound).

5 Related Work

Delta-rule state tracking. The family descends from fast-weight programmers (Schlag et al., 2021). DeltaProduct (Siems et al., 2025) characterizes which group word problems products of Householder maps solve in finite precision, proving norm stability for its own family while showing RWKV-7 unstable (Peng et al., 2025); Grazzi et al. (2025) unlock state tracking via negative eigenvalues; Yang et al. (2025) gate the recurrence; test-time-training layers rewrite state by a descent step (Sun et al., 2025)—with a quadratic inner loss one step *is* the delta-rule write of Table 1. These results address expressiveness and forward norm stability; neither addresses whether a write can be exactly *undone*—revision is an orthogonal axis, and on it we supply the construction (Theorems 1, 3, 6). Our undo step $\|A'\|_2 = 1/|1-\beta| > 1$ sits in deliberate, quantified tension with their constructive $\|A\|_2 \leq 1$: undoing requires counter-decay.

Vector-symbolic architectures. Self-inverse binding is classical in VSA—Kanerva's binary spatter codes bind and unbind by the same XOR (Kleyko et al., 2022)—but operates on feature vectors; we lift the involution to the *transition operators* of a recurrent network, where composition over time turns bit-level inverse sharing into Theorem 1.

Orthogonal RNNs. Givens/Householder parameterizations of recurrent maps are classical (Arjovsky et al., 2016; Jing et al., 2017; Mhammedi et al., 2017), maintaining a *vector* state with orthogonal maps; the operator state, the inverse-bound vocabulary, and the revisability theory are new.

Invertible networks. Reversibility has been used as a *global* architectural property—storing no activations in backpropagation (Gomez et al., 2017; Behrmann et al., 2019), memory-training RNNs (MacKay et al., 2018), or as coupling layers for density estimation (Dinh et al., 2014; 2017): the whole forward map is an engineered bijection serving optimization, with no token-level undo or inverse-closed write family. Here invertibility is local at the vocabulary level: each write ships with its exact inverse as another write, and the theorems price what that costs (Theorem 6) and buys (Theorems 1, 3).

Stacks and Dyck languages. LIFO revision is the home turf of stack-augmented networks (Joulin & Mikolov, 2015; Grefenstette et al., 2015; DuSell & Chiang, 2023; Hewitt et al., 2020; Bhattamishra et al., 2020): the stack is an $O(\text{depth})$ memory operated by learned, approximate pushes and pops. Inverse binding keeps the state constant-size and puts the LIFO discipline in the algebra itself; the deep-nesting task of Section 4.4 is a group-labelled Dyck sequence, and exactness at depth 128 comes from cancellation, not from remembering the stack.

Complexity, systems-level revision, grokking. Merrill & Sabharwal (2023); Merrill et al. (2024) and Liu et al. (2023) establish the transformer side of state-tracking complexity, Barrington (1989) the group side (Proposition 5b); knowledge-conflict surveys and KV-cache editing document and engineer the demand for revision (Xu et al., 2024; Li, 2026; Moschella et al., 2026). Grokking studies (Power et al., 2022; Stander et al., 2023) ask *when* networks learn group composition; we supply the *what*—the family property that makes the learned solution exactly revisable.

6 Conclusion

The write/undo symmetry is an architectural degree of freedom that current state-of-the-art families leave unused. Inverse-bound writes activate it: undo becomes a write, exactness becomes parameter-free (Theorem 1), drift separates from learning (Theorem 3), the DeltaNet port is priced (Theorem 6), and the taxonomy of Section 4 orders itself as the theory demands—capability follows representability of the inverse; binding makes it free. Rollback and bracketed edits are the strictly nested core; out-of-order supersession remains an addressing problem atop this primitive.

Reproducibility statement

All reported numbers are produced by the archived scripts with logged JSON outputs; theorem-relevant unit tests (21 portability + 12 K-chain + 3 nested-undo checks), the state-localization diagnostics, and all result files ship in the supplementary material. Proofs with their measured counterparts are in Appendix A; experimental protocols in Appendix B.

Use of AI tools statement

In preparing this work the authors used AI-assisted tools (an LLM-based coding agent) in the following capacities, under author direction, review, and final approval in every case: (i) aiding and polishing the manuscript's prose; (ii) retrieval and discovery of related work—every citation surfaced by the search was verified against the primary source (publisher page or arXiv abstract page) by the authors; (iii) drafting sections of the paper from author-specified content—theorem statements, proofs, and experimental protocols were specified and verified by the authors, and every numerical claim was cross-checked against archived script outputs. The conduct of the work was the authors': all experiments and unit tests were executed by the authors from the archived scripts; every reported number is script-generated, none is transcribed from model output; and each derived undo formula is backed by a unit test with a deliberately wrong variant that must fail. The scientific decisions (problem selection, claims, scope, and the negative results reported) are the authors' own.

References

Martin Arjovsky, Amar Shah, and Yoshua Bengio. Unitary evolution recurrent neural networks. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, 2016.

David A. Mix Barrington. Bounded-width polynomial-size branching programs recognize exactly those languages in NC^1. *Journal of Computer and System Sciences*, 38(1):150–164, 1989.

Jens Behrmann, Will Grathwohl, Ricky T. Q. Chen, David Duvenaud, and Jörn-Henrik Jacobsen. Invertible residual networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, 2019. arXiv:1811.00995.

Satwik Bhattamishra, Arkil Patel, and Navin Goyal. On the ability of self-attention networks to recognize counter languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7096–7116, 2020.

Max Dehn. Über unendliche diskontinuierliche gruppen. *Mathematische Annalen*, 71:116–144, 1912.

Laurent Dinh, David Krueger, and Yoshua Bengio. NICE: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014. ICLR 2015 workshop.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using Real NVP. In *International Conference on Learning Representations (ICLR)*, 2017. arXiv:1605.08803.

Brian DuSell and David Chiang. The surprising computational power of nondeterministic stack RNNs. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2210.01343.

Merrick Furst, James B. Saxe, and Michael Sipser. Parity, circuits, and the polynomial-time hierarchy. *Mathematical Systems Theory*, 17:13–27, 1984.

Aidan N. Gomez, Mengye Ren, Raquel Urtasun, and Roger B. Grosse. The reversible residual network: Backpropagation without storing activations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. arXiv:1707.04585.

Riccardo Grazzi, Julien Siems, Arber Zela, Jörg Franke, Frank Hutter, and Massimiliano Pontil. Unlocking state-tracking in linear RNNs through negative eigenvalues. In *International Conference on Learning Representations (ICLR)*, 2025.

Edward Grefenstette, Karl Moritz Hermann, Mustafa Suleyman, and Phil Blunsom. Learning to transduce with unbounded memory. In *Advances in Neural Information Processing Systems (NIPS)*, 2015. arXiv:1506.02516.

Johan Håstad. Almost optimal lower bounds for small depth circuits. In *Proceedings of the 18th Annual ACM Symposium on Theory of Computing (STOC)*, pp. 6–20, 1986.

John Hewitt, Michael Hahn, Surya Ganguli, Percy Liang, and Christopher D. Manning. RNNs can generate bounded hierarchical languages with optimal memory. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1978–2010, 2020. arXiv:2010.07515.

Nicholas J. Higham. *Accuracy and Stability of Numerical Algorithms*. SIAM, 2nd edition, 2002.

Li Jing, Yichen Shen, Bohdan Dubček, John Peurifoy, Samuel Skirlo, Yann LeCun, Max Tegmark, and Marin Soljačić. Tunable efficient unitary neural networks (EUNN) and their application to RNNs. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, 2017. arXiv:1612.05231.

Armand Joulin and Tomas Mikolov. Inferring algorithmic patterns with stack-augmented recurrent networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2015. arXiv:1503.01007.

Denis Kleyko, Dmitri A. Rachkovskij, Evgeny Osipov, and Abbas Rahimi. A survey on hyperdimensional computing aka vector symbolic architectures, Part I: Models and data transformations. *ACM Computing Surveys*, 55(6), 2022. doi: 10.1145/3538531. arXiv:2111.06077.

Bojie Li. Models take notes at prefill: KV cache can be editable and composable, 2026. arXiv:2606.17107.

Bingbin Liu, J. Trevor Ash, Surbhi Goel, Akshay Krishnamurthy, and Chi Zhang. Transformers learn shortcuts to automata. In *International Conference on Learning Representations (ICLR)*, 2023.

Matthew MacKay, Paul Vicol, Jimmy Ba, and Roger Grosse. Reversible recurrent neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

William Merrill and Ashish Sabharwal. The parallelism tradeoff: Limitations of log-precision transformers. *Transactions of the Association for Computational Linguistics*, 11:531–545, 2023.

William Merrill, Jackson Petty, and Ashish Sabharwal. The illusion of state in state-space models. In *International Conference on Machine Learning (ICML)*, 2024.

Zakaria Mhammedi, Andrew Hellicar, Ashfaqur Rahman, and James Bailey. Efficient orthogonal parametrisation of recurrent neural networks using householder reflections. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, 2017.

Luca Moschella, Laura Manduchi, and Ozan Sener. Learning to evict from key-value cache. In *International Conference on Machine Learning (ICML)*, 2026. arXiv:2602.10238.

Bo Peng, Ruichong Zhang, Daniel Goldstein, Eric Alcaide, Haowen Hou, Janna Lu, William Merrill, Guangyu Song, Kaifeng Tan, Saiteja Utpala, Nathan Wilce, Johan S Wind, Tianyi Wu, Daniel Wuttke, and Christian Zhou-Zheng. RWKV-7 "goose" with expressive dynamic state evolution. *arXiv preprint arXiv:2503.14456*, 2025.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022.

Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. Linear transformers are secretly fast weight programmers. In *International Conference on Machine Learning (ICML)*, 2021.

Julien Siems, Timur Carstensen, Arber Zela, Frank Hutter, Massimiliano Pontil, and Riccardo Grazzi. Deltaproduct: Improving state-tracking in linear RNNs via householder products. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2502.10297.

Dashiell Stander, Qihan Yu, Yuandong Lyu, Joel A. Tropp, and Joan Bruna. Grokking group multiplication with cosets. *arXiv preprint arXiv:2312.06581*, 2023.

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 Conference on Language Modeling (COLM)*, 2025. arXiv:2407.04620.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8541–8565, 2024.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving Mamba2 with delta rule. In *International Conference on Learning Representations (ICLR)*, 2025.

A PROOFS

This appendix proves the statements of Section 3 against the implementation that produced all measurements (*as-executed* proofs): every bound refers to the floating-point writes the network actually executes, not to idealized ones.

A.1 FLOATING-POINT MODEL AND AS-EXECUTED NOTATION

We use the standard model for round-to-nearest arithmetic: for $\circ \in \{+, -, \times, \div\}$, $\mathrm{fl}(a \circ b) = (a \circ b)(1+\delta)$ with $|\delta| \le \mathsf{u}$ ($\mathsf{u} = 2^{-24}$ in fp32, 2^{-53} in fp64). Length-n inner products satisfy $|\mathrm{fl}(a^\top b) - a^\top b| \le \gamma_n |a|^\top |b|$ with $\gamma_n := n\mathsf{u}/(1-n\mathsf{u})$; elementwise operations carry $O(\mathsf{u})$ relative error each (Higham, 2002). Norms are spectral (submultiplicative). Constants written c_* depend only on the operator size r, the materialization depth, and the induced bounds of Assumption H1—never on parameter values.

The write as executed. A token's rotation is a single-plane Givens update. The layer never materializes a dense matrix: it stores axes $\hat{u}, \hat{w} \in \mathbb{R}^r$ (produced by normalization and one Gram–Schmidt step from learned directions) and entries $\hat{c} := \mathrm{fl}(\cos\theta)$, $\hat{s} := \mathrm{fl}(\sin\theta)$, and applies, with $a := \hat{u}^\top X$, $b := \hat{w}^\top X$,

$$\mathrm{fl}\big(\texttt{rotate_plane}(X)\big) \;=\; X + \hat{u}\big[(\hat{c}-1)\,a - \hat{s}\,b\big] + \hat{w}\big[\hat{s}\,a + (\hat{c}-1)\,b\big]. \tag{5}$$

The *materialized rotation* of the write is the exact matrix built from the stored floats,

$$\widehat{R} \;:=\; I + (\hat{c}-1)\big(\hat{u}\hat{u}^\top + \hat{w}\hat{w}^\top\big) + \hat{s}\big(\hat{w}\hat{u}^\top - \hat{u}\hat{w}^\top\big). \tag{6}$$

All statements are about $\widehat{R}$: if $\mathrm{fl}(\cos\theta)$ is far from $\cos\theta$, the undo still cancels $\widehat{R}$, because the inverse token reuses the same stored entries.

Assumption A(i) in implementation form. The undo path never evaluates the libm at $-\theta$. The inverse token reuses the forward write's stored entries and applies the transposed rotation by flipping the sign bit of $\hat{s}$ (an exact IEEE-754 negation), so the inverse rotation is computed from the same $(\hat{c}, \hat{s})$ that parameterize $\widehat{R}$. Transposition maps equation 6 with entries $(\hat{c}, -\hat{s})$ to equation 6 with $(\hat{c}, \hat{s})$, so the materialized inverse rotation is *exactly* the transpose: $\widehat{R}(a^-) = \widehat{R}(a)^\top$ as an identity of floating-point matrices—*by construction*, with no even/odd-parity property of the libm involved. Assumption A(ii) is the implemented inverse translation $v(a^-) = \mathrm{fl}(-\widehat{R}^\top v(a))$, likewise computed from the shared entries and analyzed as a rounded matvec in the proof below.

A.2 Proof of Theorem 1 (constructive undo)

Gram–Schmidt loses orthogonality at rate $O(\mathsf{u}/\sin\angle)$ on nearly collinear inputs—the source of the $1/\Delta$ constants; anchoring the second axis to the coordinate direction least aligned with the first gives $\sin\angle \geq \sqrt{1-1/r}$ *by construction*, so the plane separation of H1 is a design choice, not a hope.

Lemma A.1 (Plane-basis deviation)**.** *The stored axes satisfy* $\big|\|\hat{u}\|_2 - 1\big| \leq c_b\mathsf{u}$, $\big|\|\hat{w}\|_2 - 1\big| \leq c_b\mathsf{u}$, *and* $|\hat{u}^\top \hat{w}| \leq c_b\mathsf{u}$, *with* $c_b = O(r/\Delta)$ *under the plane-separation clause of Assumption H1.*

Proof. The axes are produced by Euclidean normalization followed by one Gram–Schmidt projection and re-normalization, each a fixed-depth computation over length-r vectors: normalization leaves relative deviation $\leq \gamma_r$; the projection removes the $\hat{u}$-component of d_2 up to an inner-product error $\leq \gamma_r \|d_2\|_2$ plus update rounding of the same order, leaving a residual $\hat{u}$-component of that size in w_{perp}. By plane separation the exact perpendicular component has norm $\geq \sin\angle(d_1, d_2)\, \|d_2\|_2 \geq \Delta\, \|d_2\|_2$, so the final normalization amplifies the residual to $|\hat{u}^\top \hat{w}| \leq c\,\gamma_r/\Delta$; the remaining steps give constants linear in r/Δ. □

Lemma A.2 (Pair-orthogonality deviation)**.** *For every stored* $(\hat{u}, \hat{w}, \hat{c}, \hat{s})$ *satisfying Lemma A.1 and Assumption H1,* $\left\|\widehat{R}^\top \widehat{R} - I\right\|_2 \leq c_0\mathsf{u}$ *with* $c_0 = O(c_b)$*; in particular* $\left\|\widehat{R}\right\|_2 \leq 1 + c_0\mathsf{u}$.

Proof. Let $U := [\hat{u}\ \hat{w}]$ and $J := \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$, so that equation 6 reads $\widehat{R} = I + (\hat{c}-1)UU^\top - \hat{s}\,UJU^\top$. Decompose $\mathbb{R}^r = \mathrm{span}(U) \oplus \mathrm{span}(U)^\perp$. On the span, $U^\top U = I_2 + E_0$ with $\|E_0\| \leq c_b\mathsf{u}$ by Lemma A.1, so the restriction of $\widehat{R}^\top \widehat{R}$ is $(\hat{c}^2 + \hat{s}^2)\, I_2$ up to skew mixing of size $O(c_b\mathsf{u})$; the cross terms against E_0 and the leakage into the complement are $O(c_b\mathsf{u})$ of the same origin. For the leading term, $\hat{c} = \cos\bar{\theta}\,(1+\delta_1)$ and $\hat{s} = \sin\bar{\theta}\,(1+\delta_2)$ with $|\delta_i| \leq \mathsf{u}$ (rounding of the transcendental evaluation), hence $(1-\mathsf{u})^2 \leq \hat{c}^2 + \hat{s}^2 \leq (1+\mathsf{u})^2$ and $|\hat{c}^2 + \hat{s}^2 - 1| \leq 2\mathsf{u} + \mathsf{u}^2$. Note the squares are exact: $\widehat{R}$ is an exact matrix built from stored floats. Collecting terms gives $c_0 = O(c_b)$. The norm bound follows from $\left\|\widehat{R}\right\|_2^2 \leq \left\|\widehat{R}^\top \widehat{R}\right\|_2 \leq 1 + c_0\mathsf{u}$. □

Remark A.3 (The leading term measured at zero). In every logged run, $\hat{c}^2 + \hat{s}^2$ rounds *exactly* back to 1 (both precisions, all tested θ), so the $2\mathsf{u}$ term of Lemma A.2 is measured at 0. This is an empirical property of the libm evaluations, not an assumed one; the theorem does not rely on it, and the K-chain measurements (Appendix C) show its consequence: the linear-in-K accumulation channel is absent at its source.

Lemma A.4 (Single-write application error)**.** *For every stored write and every* $X \in \mathbb{R}^{r\times r}$, $\mathrm{fl}\big(\texttt{rotate_plane}(X)\big) = \widehat{R}\,X + E$ *with* $\|E\|_2 \leq c_e\mathsf{u}\, \|X\|_2$, *where* $c_e = O(r)$ *depends only on* r *and the materialization depth.*

Proof. The two projections a, b are r-length inner products (error $\gamma_r \|X\|_2$ in the rows they touch); the update equation 5 adds to two rows a linear combination with coefficients of modulus ≤ 2 (Assumption H1: $|\hat{c}|, |\hat{s}| \leq 1$), each operation contributing $O(\mathsf{u})$ relative error on quantities already bounded by $2\,\|X\|_2$ up to Lemma A.1 deviations. Summing the per-element bounds row-wise and taking norms gives $c_e = O(r)$. □

Proof of Theorem 1. P component. The executed forward step gives $P_1 = \mathrm{fl}(\widehat{R}P) = \widehat{R}P + E_1$ with $\|E_1\|_2 \leq c_e\mathsf{u}\, \|P\|_2$ (Lemma A.4). The executed undo step gives $\widetilde{P} = \mathrm{fl}(\widehat{R}^\top P_1) + E_2'$, i.e.,

$$\widetilde{P} \;=\; \widehat{R}^\top \widehat{R}\, P + \widehat{R}^\top E_1 + E_2 \;=\; P + (\widehat{R}^\top \widehat{R} - I)P + \widehat{R}^\top E_1 + E_2,$$

using $\widehat{R}(a^-) = \widehat{R}(a)^\top$ exactly (Section A.1), with $\|E_2\|_2 \leq c_e\mathsf{u}\, \|P_1\|_2 \leq c_e\mathsf{u}\,(1+c_0\mathsf{u}+c_e\mathsf{u})\, \|P\|_2$. Taking norms with Lemma A.2 ($\left\|\widehat{R}^\top \widehat{R} - I\right\|_2 \leq c_0\mathsf{u}$, $\left\|\widehat{R}\right\|_2 \leq 1 + c_0\mathsf{u}$):

$$\left\|\widetilde{P} - P\right\|_2 \;\leq\; \big[c_0 + 2c_e(1 + c_0\mathsf{u}) + c_e^2\mathsf{u}\big]\, \mathsf{u}\, \|P\|_2 \;=:\; c_1\mathsf{u}\, \|P\|_2\,.$$

c component. The forward step yields $c_1 = \widehat{R}c + v + E_3$, $\|E_3\|_2 \le c_e\mathsf{u}\,(\|c\|_2 + \|v\|_2)$ (two-vector affine combination, same analysis as Lemma A.4). The inverse translation is $v^- = \mathrm{fl}(-\widehat{R}^\top v) = -\widehat{R}^\top v + E_v$, $\|E_v\|_2 \le c_e\mathsf{u}\ \|v\|_2$ (rounded matvec). The undo step yields

$$\widetilde{c} \;=\; \widehat{R}^\top c_1 + v^- + E_4 \;=\; c + (\widehat{R}^\top\widehat{R} - I)\,c + \widehat{R}^\top E_3 + E_v + E_4,$$

with $\|E_4\|_2 \le c_e\mathsf{u}\,(\|c_1\|_2 + \|v\|_2)$. Collecting, $\|\widetilde{c} - c\|_2 \le c_2\mathsf{u}\,(\|c\|_2 + \|v\|_2)$ with $c_2 = c_0 + 4c_e(1 + c_0\mathsf{u})(1 + c_e\mathsf{u}) = O(c_e)$. All constants depend on r, materialization depth, and H1 only—not on parameter values. □

Remark A.5 (Reachable states make the bound absolute). OS states start at $(I, 0)$ and evolve as products of materialized rotations, so $\|P\|_2 \le (1 + c_0\mathsf{u})^t$ after t steps. For any realistic horizon ($t\mathsf{u} \ll 1$) this factor is $1 + O(t\mathsf{u})$, and the relative statement of Theorem 1 is an absolute $O(\mathsf{u})$ one. The relative form is what the proof yields; we state it as such.

A.3 Proof of Proposition 2 (unbound dual)

Proof of Proposition 2. Work in homogeneous coordinates, with $M(a)$ the exact $(r+1) \times (r+1)$ materialized operator of token a (invertible: its rotation block is $\widehat{R}$ with $\left\|\widehat{R}\right\|_2 \ge 1/(1+c_0\mathsf{u}) > 0$ by Lemma A.2) and $\widehat{M}$ the materialized operator of the independently parameterized inverse symbol. Both are exact matrices over the reals, and

$$\widehat{M}\,M(a) - I \;=\; \big(\widehat{M} - M(a)^{-1}\big)M(a)$$

is an identity of real matrices—no floating-point quantity enters. The executed composition differs from this product by the application errors of Lemma A.4 ($O(\mathsf{u})$, the same terms present under binding), so the undo error is the learned mismatch $(\widehat{M} - M(a)^{-1})M(a)$ up to an additive $O(\mathsf{u})$; it vanishes exactly when the learned inverse is exact. □

A.4 Proof of Theorem 3 (drift separation)

Proof of Theorem 3(a). Each block satisfies, exactly, $\widehat{M}_k M_k = M_k^{-1}M_k + (\widehat{M}_k - M_k^{-1})M_k = I + F_k$ with $F_k := (\widehat{M}_k - M_k^{-1})M_k$. Substituting blockwise, $\prod_{k=1}^K \widehat{M}_k M_k = \prod_{k=1}^K (I + F_k)$ as an identity of real matrices, for every realization of the mismatches. Under binding, Theorem 1 gives $\|F_k\|_2 = O(\mathsf{u})$ (plus the application errors, likewise $O(\mathsf{u})$). □

Lemma A.6 (Repeated mismatch accumulates at first order)**.** *Let $F_1 = \cdots = F_K = F$ with $\delta := \|F\|_2$ and $K\delta \le \frac{1}{2}$. Then*

$$\big\|(I + F)^K - I\big\|_2 \;\ge\; K\delta\big(1 - K\delta\big).$$

Proof. Expand $(I + F)^K - I = \sum_{j\ge1}\binom{K}{j}F^j$. The first-order term has norm $K\delta$; the remainder has norm at most $\sum_{j\ge2}\binom{K}{j}\delta^j = (1 + \delta)^K - 1 - K\delta$. By the reverse triangle inequality,

$$\big\|(I + F)^K - I\big\|_2 \;\ge\; K\delta - \big[(1 + \delta)^K - 1 - K\delta\big].$$

Now $(1 + \delta)^K \le e^{K\delta}$, and $e^x - 1 - x \le \frac{x^2}{2}e^x$ for $x \ge 0$ gives $(1 + \delta)^K - 1 - K\delta \le \frac{e^{1/2}}{2}(K\delta)^2 \le (K\delta)^2$ when $K\delta \le \frac{1}{2}$. Cancellation cannot occur at first order: the first-order term is KF, the same F repeated K times. □

Proof of Theorem 3(b) and (c). Part (b) is Lemma A.6 restated. For (c): under binding, each adjacent pair composes to $I + E_k$ with $\|E_k\|_2 \le c_1\mathsf{u}$ (Theorem 1), so $P_K = \prod_k (I + E_k)\,P_0$ and

$$\|P_K - P_0\|_2 \;\le\; \Big[\prod_k (1 + c_1\mathsf{u}) - 1\Big]\,\|P_0\|_2 = \big[(1 + c_1\mathsf{u})^K - 1\big]\,\|P_0\|_2 \;\le\; Kc_1\mathsf{u}\,e^{Kc_1\mathsf{u}}\,\|P_0\|_2\,,$$

using $(1 + x)^K \le e^{Kx}$ and $e^x - 1 \le xe^x$. The c component telescopes the same way over K pairs: each pair contributes $\le c_2\mathsf{u}(\|c_k\|_2 + \|v_k\|_2)$ (Theorem 1) and the intermediate rotations amplify by at most $(1 + c_0\mathsf{u})^K$, giving $\|c_K - c_0\|_2 \le Kc_2\mathsf{u}\max_k(\|c_k\|_2 + \|v_k\|_2)(1 + c_0\mathsf{u})^K$. The constants depend on machine precision, K, and data scale only. □

A.5 Proof of Theorem 6 (portability)

Lemma A.7 (Sherman–Morrison inverse, general form)**.** *Let $A = I - \beta kk^\top$ with $s := \|k\|_2^2$ and $\beta s \neq 1$. Then A is invertible with*

$$A^{-1} \;=\; I + \tfrac{\beta}{1-\beta s}\, kk^\top \;=\; I - \beta' kk^\top, \qquad \beta' := -\tfrac{\beta}{1-\beta s}.$$

Proof. Direct verification: $AA^{-1} = I + \frac{\beta - \beta(1-\beta s) - \beta^2 s}{1-\beta s}\, kk^\top = I$, since the numerator vanishes identically. A is a rank-one modification of I with eigenvalue 1 on $k^\perp$ and $1-\beta s$ along k, so $\beta s = 1$ is exactly the non-invertible face. □

Proof of Theorem 6(a) and (b). (a) The undo write uses (k, v, β') with $\beta' = -\beta/(1-\beta s)$; at $\|k\|_2 = 1$, $s = 1$ and this is equation 4. Linear part: $A'A = I$ by Lemma A.7. Affine part: $A'k = (1-\beta' s)k$ and $1 - \beta' s = \frac{1}{1-\beta s}$, hence $A'B = \beta\,(A'k)\,v^\top = \frac{\beta}{1-\beta s}\, kv^\top = -\beta' kv^\top = -B'$, so the composite map is $S \mapsto A'(AS_0 + B) + B' = S_0$ for *all* S_0: undo is one write of the same family. *(b)* By Lemma A.7 the only non-invertible face is $\beta s = 1$. At $\|k\|_2 = 1$, $\beta = 1$: $A = I - kk^\top$ is the orthogonal projection onto $k^\perp$, and $S \mapsto AS + kv^\top$ replaces the k-component of S by the constant $v^\top$: two states differing only along k map to the same output (measured: $\|\Delta_{\text{in}}\|_2 = 5.13 \mapsto \|\Delta_{\text{out}}\|_2 = 2.1 \times 10^{-15}$). Off this face the family is invertible on an open dense set. □

Lemma A.8 (Finite-precision undo)**.** *Executing the write and undo of Theorem 6(a) in floating point over a d-dimensional state satisfies*

$$\left\| \widetilde{S} - S_0 \right\|_2 \;\le\; c\,\gamma_d\,\kappa\,\big(\|S_0\|_2 + \|v\|_2\big), \qquad \kappa := \|A'\|_2 = \max\big(1,\ |1-\beta s|^{-1}\big),$$

with c a small absolute constant. The degeneration toward the singular face is $O(\mathsf{u}/|1-\beta s|)$, exactly as priced in Theorem 6(c).

Proof. The forward step executes as $S_1 = AS_0 + B + E_1$ with $\|E_1\|_2 \le c\,\gamma_d(\|S_0\|_2 + \|v\|_2)$ (rank-one update and rescaling over d-wide vectors). The undo executes as $\widetilde{S} = A'S_1 + B' + E_2$ with $\|E_2\|_2 \le c\,\gamma_d(\kappa\,\|S_0\|_2 + \|v\|_2)$. By the exact algebra of Theorem 6(a), all algebraic terms cancel and $\widetilde{S} = S_0 + A'E_1 + E_2$; conclude with $\|A'\|_2 = \kappa$. □

Remark A.9 (Nested undo compounds the port's price). A bracketed DeltaNet undo of depth d executes d undo writes in reverse order. In exact arithmetic the intermediate states are exactly the historical states (each undo is the algebraic inverse of a write), so the state does not blow up—but each step's error (Lemma A.8) is amplified by the product of the remaining undo norms, each $\le \kappa = \|A'\|_2$, giving total error $O(d\,\kappa^d \mathsf{u})$: exponential in depth. The orthogonal carrier has $\kappa = 1$ at every step and pays $O(d\mathsf{u})$. Measured (`diag_nested_undo`; fixed (k, v, β) chain, the recurring regime of Theorem 3(b); mean over 3 seeds): fitted $\log_2$-slopes of the depth-d undo error are 0.99 ($\beta = \frac{1}{2}$, $\kappa = 2$) and 1.57 ($\beta = \frac{2}{3}$, $\kappa = 3$) against predictions $\log_2 \kappa = 1$ and 1.58; at depth 64 (fp32) the error reaches 1.7×10^{11} and 3.6×10^{22}, while the orthogonal chain of the same depths stays at 3.0×10^{-6} (slope 0.02; flat through depth 128 in fp64). Random per-step keys (typical regime) damp the base but stay exponential: slope 0.11 at $\kappa = 2$, r=16.

Proof of Theorem 6(c). (i) Domain. The forward parameterization is $\beta = 2\,\sigma(z) \in (0, 2)$ with σ the logistic function (verified in the reference implementation; the range is open). For $\beta \in (0, 1)$: $\beta' = -\beta/(1-\beta) < 0$. For $\beta \in (1, 2)$: $\beta' = \beta/(\beta - 1) > 2$, since $\beta/(\beta-1) > 2 \iff \beta < 2$. Hence $\beta' \notin (0, 2)$ for every forward-admissible β; the only value mapping into the domain is the excluded endpoint $\beta = 2 \mapsto \beta' = 2$. As β' lies outside the range of 2σ, the undo write cannot be produced by any parameter value or clamp: it must be injected structurally. *(ii) Expansion.* By Lemma A.7, A' has eigenvalue $1 - \beta' s = \frac{1}{1-\beta s}$ along k; at $s = 1$ this exceeds 2 in modulus iff $|1-\beta| < \frac{1}{2}$, and diverges as $\beta \to 1$. *(iii) Stability tension.* A' is symmetric with spectrum $\{1^{(d-1)}, \frac{1}{1-\beta s}\}$, so $\|A'\|_2 = \max(1, |1-\beta s|^{-1}) > 1$ for every $\beta s \in (0, 2)$, $\beta s \neq 1$. The forward family satisfies $\|A\|_2 = \max(1, |1-\beta s|) \le 1$ constructively on the same domain, and this boundedness is precisely the forward stability the family is designed for; an exact undo necessarily exits

the stable set (counter-decay). The orthogonal family undoes with $\left\|R^{-1}\right\|_2 = 1$ exactly: inversion is an isometry there. □

Proof of Theorem 6(e). One step interleaves n_h sub-writes as $A = A_{n_h} \cdots A_1$ (reference implementation), so its inverse is $A^{-1} = A_1^{-1} \cdots A_{n_h}^{-1}$: the undo sub-writes must execute in reverse order, each an instance of Theorem 6(a) on its own (k_i, β_i). Invertibility itself is order-free; only the execution order of the undo is constrained. Measured: reverse order 1.08×10^{-14} vs. forward order, separated by 5.5×10^{13} (fp64). □

Proof of Corollary 7. Each in-domain write $H_i = I - \beta_i k_i k_i^\top$ is symmetric with spectrum $\{1^{(d-1)}, 1-\beta_i\}$ and $|1-\beta_i| < 1$, so $\|H_i\|_2 = 1$: every write is non-expansive and strictly contractive along k_i. Suppose $W'W = I$ with both words in-domain, $W = H_n \cdots H_1$. For every x, $\|x\|_2 = \|W'Wx\|_2 \le \|Wx\|_2 \le \|x\|_2$, so W is an isometry, hence orthogonal; in particular invertible, so $\det W = \prod_i (1-\beta_i) \neq 0$ and every prefix $H_j \cdots H_1$ is surjective. Chaining the non-expansive factors, $\|x\|_2 = \|Wx\|_2 \le \|H_{n-1} \cdots H_1 x\|_2 \le \cdots \le \|x\|_2$ forces equality throughout; the j-th link reads $\|H_j y\|_2 = \|y\|_2$ for all y in the image of $H_{j-1} \cdots H_1$, which is the full space. So each H_j is an isometry, and $H_j^\top H_j = H_j^2 = I$ iff $\beta_j(\beta_j - 2)\, k_j k_j^\top = 0$ iff $\beta_j \in \{0, 2\}$—both outside the open range $2\sigma(z) \in (0, 2)$. □

Part (d) is proved in Appendix D, which also records the kernel convention from which the uniqueness statement is read off.

A.6 Proof–measurement correspondence

Table 3 pairs each proved statement with its measured counterpart, and Table 4 details the portability unit tests cited in Section 4.5; all files are archived with logged outputs.

Table 3: Statements and their measured counterparts (`results/*.json`).

Statement	Assertion	Measured
Thm. 1	one round $O(\mathsf{u})\,\|P\|_2$	$2\mathsf{u} = 1.19\text{e}{-7}$ (fp32); $2.2/4.4\text{e}{-16}$ (fp64)
Prop. 2	residual = learned mismatch	direct (Fig. 3): $1.2\text{e}{-7}$ (bound) vs $\sim 2\text{e}{-1}$
Thm. 3(c)	K-rounds $\le KO(\mathsf{u})$	K=1024: $7.15\text{e}{-6} \ll 50K\mathsf{u} = 3.1\text{e}{-3}$ (fp32)
Thm. 6(a)+Lem. A.8	exact undo; $O(\mathsf{u}/\lvert 1-\beta\rvert)$	$1.15\text{e}{-15}/8.29\text{e}{-7}$; slopes $1.00/0.99$ (fp64)
Thm. 6(b)	$\beta s = 1$ irreversible	$5.13 \mapsto 2.06\text{e}{-15}$
Thm. 6(e)	reverse order required	$1.08\text{e}{-14}$; forward separated $\times 5.5\text{e}13$
Thm. 6(d)	unique gated undo	$1.24/1.78\text{e}{-16}$; in-domain fails at 0.89
Rem. A.9	nested DN undo $\sim \kappa^d$; OS flat	slopes $0.99/1.57$; d=64: $1.7\text{e}11$ vs $3.0\text{e}{-6}$

Table 4: Portability tests (`results/port_results.json`). "Separation" = error ratio of the wrong variant to the correct formula. The gated row lists two fp64 sub-cases—the no-forget-gate configuration ($\alpha \equiv 1$) and the general gated undo equation 7; fp32: 7.41×10^{-8}, 1.67×10^{-7}.

Test	Formula	Error (fp64 / fp32)	Wrong variant
Unit-key undo	equation 4	$1.15\text{e}{-15}$ / $8.29\text{e}{-7}$	$\beta' = -\beta$: $\times 2.5\text{e}13$
General $\|k\|_2 \neq 1$	general form	$7.99\text{e}{-15}$ / $3.89\text{e}{-6}$	unit form: fails (2.18)
$\beta \to 1$ scaling	$O(\mathsf{u}/\lvert 1-\beta\rvert)$	slopes $1.00/0.99$ (fp64)	β=1: irreversible
n_h=2 order	reverse order	$1.08\text{e}{-14}$	forward order: $\times 5.5\text{e}13$
Gated (α=1; general)	equation 7	$1.24\text{e}{-16}$, $1.78\text{e}{-16}$ (fp64)	in-domain $\alpha' = \alpha$: 0.89

B Experimental details

Hardware and precision. All training runs on a single RTX 3060 Laptop GPU (6 GB, PyTorch 2.5.1+cu121) in fp32; all algebraic verification (unit tests, scan identities, portability and K-chain measurements) additionally in fp64 on CPU. We deliberately do not train OS in mixed precision:

the exactness claims are fp32 claims, and the training runs themselves are their evidence. Registered deviations from reference-implementation defaults (batch size, curriculum) are noted per task below.

Program-evaluation tasks (Section 4.1). Programs are fixed-length token sequences over a generator vocabulary with formal inverses; the label is the ordered product evaluated stepwise, read out at a final query token. *The task group is concrete*: the 24 operations are independent random 3-cycles on 12 points (drawn once, seed 1234; a 3-cycle's inverse is again a 3-cycle, and every operation is an even permutation). The subgroup they generate was certified by Schreier–Sims: it has order $239{,}500{,}800 = |A_{12}|$ exactly—the full alternating group, which is nonsolvable, so the Barrington backdrop of Proposition 5(b) applies to the task. Every generator is single-plane-representable by construction: a 3-cycle is a 120° rotation in a single plane (within its support, about the $(1,1,1)$ axis). This is a vocabulary-level constraint, not an architectural one—a 5-cycle spans two rotation planes and would occupy two tokens, since products of single-plane rotations over arbitrary planes generate all of $SO(r)$; the same single-plane face appears for DeltaNet at $n_h{=}1$ (Table 1). The data generator is unit-tested to guarantee: support elements distinct; forward/inverse relations are mutual inverses; program truth equals stepwise composition; a symbolic reference agrees 100%; the main split contains no adjacent inverse pairs (so cancellation is never trivially learnable), while the cancellation variants contain them by construction and their true value is computed under exact cancellation. Training streams freshly sampled programs (infinite stream, every batch new); evaluation uses a fixed held-out suite of 512 programs per length drawn from the same generator at seed 999—at $L{=}8$ the program space has $48^8 \approx 2.8 \times 10^{13}$ elements, so train/eval collisions are negligible and memorization is not a viable strategy. (The original generating script is named `diag_overfit` for historical reasons; Table 5 reports its multi-seed rerun `diag_controls`, parallel scan path, fp32 mixer equivalence 4.1×10^{-8} vs. serial verified.) Architectures (all at the matched schedule): OS $d{=}64$, $L{=}2$, $H{=}2$, $r{=}16$, ffn 256 (86,084 parameters); *same-plane* control (77,764; rotation planes fixed to one coordinate plane, so the channel is a single commuting rotation); transformer (174,336; with a matched position table $\max_T{=}16$: 104,960, scores indistinguishable—position-table size does not matter here). A third control, *zero-angle* (86,084 with W_θ frozen at zero), removes program content from the state entirely; its predictor reduces to the query-conditional marginal and is deterministic across seeds. All Table 5 cells are means over three seeds $(42/43/44)$. Training: AdamW (wd 0.01), gradient clip 1.0, $B{=}32$, 2000 steps per length, lr $1\mathrm{e}{-3}$ (kept from the historical sweep over $\{3\mathrm{e}{-4}, 1\mathrm{e}{-3}\}$; grokking plateaus make short sweeps non-discriminative).

Table 5: Program-evaluation detail: lengths $L \in \{2, 4, 8\}$ and OS-channel controls at the $1\times$ schedule (exact accuracy, mean over 3 seeds; $x \to y$ = same architecture at $10\times$, 20,000 steps; `diag_controls`, `diag_sched10x`). *same-plane* = one fixed coordinate plane (commuting channel); *zero-angle* = $\theta \equiv 0$ (query-conditional marginal). Matched position tables change nothing ($0.905/0.653/0.195$; $0.195 \to 0.317$ at $10\times$).

Length L	OS (bound)	DeltaProduct ($n_h{=}2$)	same-plane	zero-angle	Transformer
2	**1.000**	0.954	0.671	0.602	0.898
4	**1.000**	0.942	0.429	0.363	0.645
8	**1.000**	$0.390 \to 0.993$	0.201	0.181	$0.201 \to 0.320$

Table 6: Binding ablation at equal parameters and the $1\times$ schedule (87,620 both; `diag_binding_eq`; mean$\pm$std over 3 seeds; $L{=}8$). "Pure chain" = concatenated $(a\, a^-)$ pairs, true value I throughout. The $10\times$ columns of Table 2 close all three task cells; the state residual does not close (Figure 3).

Setting ($L{=}8$)	bound	unbound
main (mixed programs)	$\mathbf{1.000} \pm 0.000$	0.885 ± 0.082
cancel-heavy	$\mathbf{1.000} \pm 0.000$	0.927 ± 0.052
pure cancellation chains	$\mathbf{1.000} \pm 0.000$	0.973 ± 0.022

Extrapolation protocol (Section 4.2). Train only at $L{=}16$ (grok transition at step 7250); evaluate zero-shot at $L \in \{16, 32, 64, 128, 256, 512, 1024\}$, 256 fresh programs per length, no adjacent

inverse pairs. No positional encoding in the OS model; nothing in the training distribution contains lengths > 16.

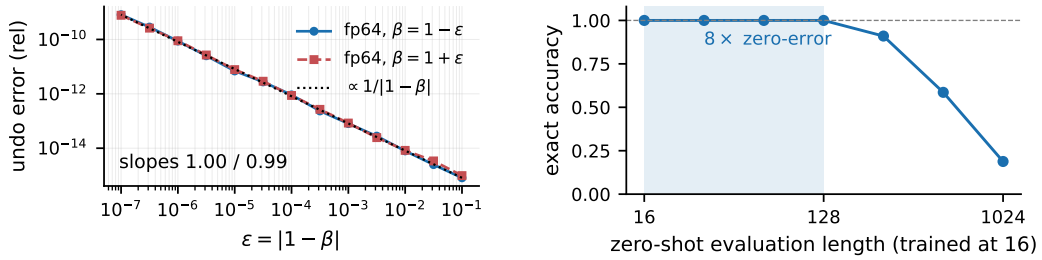


Figure 4: **Left:** Theorem 6(a): undo error tracks $O(\mathrm{u}/|1-\beta|)$, fitted slopes $1.00/0.99$ (fp64)—the degeneration is the singular face $\beta \to 1$. **Right:** zero-shot extrapolation of an L=16-trained bound model: zero error through $8\times$ training length, then linear-law decay—a property of the learned forward operators, not of the undo guarantee.

Binding ablation (Section 4.3). Identical architecture, data stream, and budget as Section 4.1, at *strictly equal parameter count* (87,620): both variants use a 60-row embedding table; the bound variant indexes only the 36 bound rows (24 dead rows). The single variable is vocabulary semantics: *bound* (each generator shares its embedding with a designated inverse symbol plus an inverse mask) vs. *unbound* (inverse actions are independent learned symbols). Three seeds; the residual probe (Figure 3) runs the same protocol, measuring $\|P_2 - I\|$ on all 48 generator/direction pairs every 100 steps. The pure-chain variant concatenates $(a\, a^-)$ pairs so the true value is I throughout; the cancel-heavy variant mixes cancellation pairs into general programs.

State-localization diagnostics (Section 4.3, Figure 2; `diag_localize`). The three 20k deep-nest models (bound, unbound, DeltaProduct) are retrained from the archived protocol (same seeds, same data streams; end-of-training residuals and task accuracies reproduce the published values), and probed without further training. Per-head pair residuals: the 48-pair probe of the residual protocol, decomposed by head and direction. State error vs. depth: on depth-d programs, $\left\|A_{\mathrm{end}} - A_{\mathrm{core}}\right\|_{\max}$ per layer (A_{core} = state after the core token; DeltaProduct analogously on S), plus the readout-visible component $\left\|(A_{\mathrm{end}} - A_{\mathrm{core}})\, h\right\|_{\max}$ at the final program position ($h = q + z$, the state-to-signal contraction). Linear probe: logistic regression on the first layer's final-position signal y_0 (concatenated heads, 32 dims) and on the raw state P_0 (flattened, 512 dims), trained on 512 depth-8 programs, tested on 256 held-out and 256 depth-128 programs. The capture path is verified bitwise-identical to the model's own recorded states.

$10\times$-schedule controls (Sections 4.1–4.4; `diag_sched10x`). Transformers (both position configurations) and the equal-parameter OS pair are retrained for 20,000 steps at L=8 (identical data stream, optimizer, seeds $\{42, 43, 44\}$; checkpoints at 2k/10k/20k). DeltaProduct baseline: reference equations of Siems et al. (2025), two Householder heads per token with negative eigenvalues ($\beta \in (0, 2)$ via 2σ), interleaved-sequence implementation (`battle_models`), d=64, H=2, head-dim 16, intermediate 256, plain 60-symbol vocabulary; 65,744 parameters (fewest of all contestants). Plain runs at $L \in \{2, 4\}$ use 2000 steps, at L=8 also 20,000; deep-nest runs use the mixed protocol above. Undo-residual curves extended to 20k steps: `diag_residual_20k`. A second $10\times$ control retrains the deep-nest mixed protocol itself for 20,000 steps (OS bound/unbound, DeltaProduct; seeds $\{42, 43, 44\}$; end-of-training residual probe on all 48 pairs): `diag_deepnest_20k`.

β-saturation check (Corollary 7; `diag_beta_check`). In fp32, $2\sigma(z)$ first rounds to exactly 2 at $z = 16.64$. Over the mixed evaluation suites (plain and cancel-heavy L=8, deep-nest $d \in \{8, 128\}$; 150,528 write evaluations per seed), the archived DeltaProduct checkpoints (`weights/`) execute $\max \beta = 1.9970/1.9970/1.9958$ (seeds 42/43/44) with $\max |z| \leq 6.5$: interior as-executed, yet pressed toward the excluded endpoint. Residual tail fits (Section 4.3, Figure 3) are produced by `fit_residual_tail`.

Deep-nested undo protocol (Section 4.4). Programs are core, $a_1 \cdots a_d$, $a_d^- \cdots a_1^-$ (a LIFO undo stack of depth d; true value = core). Training: length-8 programs, each batch alternating between plain programs and cancel-pair insertions (cancellation present but shallow); bound/unbound $\times$ seeds $\{42, 43, 44\}$; other hyperparameters as above. Evaluation: depths $\{2, 4, 8, 16, 32, 64, 128\}$ (256 programs each), with plain programs of the same total length $(2d{+}1)$ as the paired control. Generator verified: composed program equals the core permutation on 200 random instances; token/mask consistency unit-tested. Data: `diag_deepnest`.

undo-DeltaNet protocol (Section 4.5). Reference DeltaNet (nh $=$ 1, negative eigenvalues) as implemented in `battle_models` from the equations of Siems et al. (2025) (their state-tracking configuration: no forget gate), d=64, H=2, head_dim 16 (state $H \cdot d_h^2 = 512$, capacity-matched to the OS configuration $H \cdot r^2$), intermediate 256, fp32, otherwise the binding-ablation protocol. The bound variant reuses the forward token's projection (same k, v, β) and overrides $\beta \mapsto -\beta/(1-\beta)$ at inverse tokens via an undo mask; the plain variant uses the 60-symbol vocabulary as-is. Mechanism check (`b6`): an adjacent do/undo pair returns the state to $\|\Delta S\| = 2.4 \times 10^{-7}$; $\min|1-\beta| \in [0.40, 0.53]$ across runs (numerically safe regime). Data: `diag_undo_dn`.

$A_5 \subset SO(3)$ construction (Proposition 5a). The icosahedral rotation group is generated by a 72° rotation about a vertex axis and a 180° rotation about an edge-midpoint axis; BFS closure gives exactly 60 elements ($a^5 = b^2 = (ab)^3$). Each element decomposes by axis–angle (via atan2, well-conditioned at $\theta{=}\pi$) into a single-plane write (u, w, θ) with the plane orthogonal to the axis: $\max \|\texttt{rotate_plane}(X) - RX\| = 1.3 \times 10^{-15}$ over all 60 elements and a 60-step random generator word (2.2×10^{-15}), fp64 (`test_a5_so3`).

Byte-level LM and dose–response (Section 4.6). Corpus: Python 3.11 standard-library sources (PSF license), document-level split $80/10/10$, 9,356,420 training bytes, file manifest fixed by sha256. No inverse designations exist on the byte vocabulary; inverse binding is never triggered in these runs (Section 4.6). Training: B=16, T=128, 977 steps = 2,004,224 predicted bytes, every byte presented exactly once (no repetition); AdamW as above; seeds $\{42, 43, 44\}$ (dose–response means$\pm$sd in Section 4.6); learning rate selected by a 300-step sweep over $\{3\text{e}{-4}, 1\text{e}{-3}\}$ per model ($1\text{e}{-3}$ selected for all). Architectures: OS d=128, L=4, H=4, r=16 (860,432 parameters); transformer, same width/depth (957,184); hybrid = [Attn, OS, Attn, OS] (843,272). Metric: bits per byte on held-out documents; incremental decode at B=1: 281 tok/s (OS) vs. 779 (transformer).

Parallel scan. The associative form computes prefix products $P_t = R_t \cdots R_1$ by log-doubling, uses orthogonality ($P_s^{-1} = P_s^\top$) to turn the translation recurrence into $c_t = P_t \sum_{s \le t} P_s^\top v_s$, and reads out with one batched product. Verified against the incremental form: identity $\le 7.6 \times 10^{-17}$ (fp64, all mode switches $\times$ inverse-mask settings), mixer-level equivalence 4.1×10^{-8} (fp32), record-state equivalence 1.4×10^{-15}, full-model logits 2.0×10^{-7}, gradient equivalence 2.5×10^{-6} (relative); end-to-end speedup 12–13$\times$ (forward+backward).

Unit tests. Portability (`test_port`, $21/21$): each acceptance criterion is paired with a deliberately wrong variant (wrong sign, unit-key form on general keys, forward sub-step order, in-domain gate attempt) that must fail, so a wrong formula cannot pass as a small number; executed on CPU against the reference recurrence (identical equations to the released DeltaNet/DeltaProduct forward path, no compiled kernels). K-chain (`test_kchain`, $12/12$): $K \in \{1, \ldots, 1024\}$, fp32/fp64, same-pair and random-pair modes, all under $50K\,\mu$.

Provenance. Every number in the paper is produced programmatically from the archived result files by the figure/table scripts; no number is transcribed by hand.

C K-CHAIN DIAGNOSTICS: SUBLINEAR SATURATION

Setup: K adjacent (a, a^-) rounds through the exact inverse-bound code path used in training, measuring $\|P_K - I\|_2$, for $K \in \{1, 2, 4, \ldots, 1024\}$, in fp32 and fp64, with either the *same* pair repeated K times (the worst-case systematic regime of Theorem 3(b)) or a fresh *random* pair each round. All

twelve measurement series satisfy $\|P_K - I\|_2 \le 50\,K\mathrm{u}$ at every K (the test's acceptance criterion; the theorem itself needs only the $K\mathrm{u}$-rate envelope).

Table 7: K-chain measurements (`results/kchain_results.json`). Slopes are log–log fits of $\|P_K - I\|_2$ against K; "saturated at" gives the plateau value and the K at which it is reached.

Series	Slope	$\|P_{1024} - I\|_2$	Saturated at
fp32, same pair	0.62	7.15×10^{-6} (1.2×10^{2}u)	7.15×10^{-6} from K=128
fp32, random pairs	0.48	3.22×10^{-6}	not yet
fp64, same pair	0.67	1.80×10^{-14} (1.6×10^{2}u)	1.80×10^{-14} from K=256
fp64, random pairs	0.52	7.66×10^{-15}	not yet

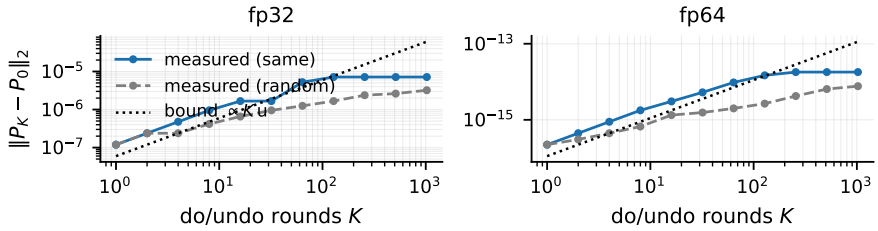


Figure 5: Constructive undo across K do/undo rounds. Measured error (solid) grows sublinearly and saturates; the parameter-free bound of Theorem 3(c) (dashed, the $K\mathrm{u}$ line) holds at every K with an order of magnitude to spare.

Two measured mechanisms explain the gap to the $K\mathrm{u}$ bound; neither is claimed by any theorem:

1. **The linear channel is absent at its source.** $\mathrm{fl}(\cos\theta)^2 + \mathrm{fl}(\sin\theta)^2$ evaluates to exactly 1 for every tested θ in both precisions (Remark A.3), so the leading 2u term of Lemma A.2 is zero and the per-round increments are higher-order from the start.
2. **fp32 reaches a rounding fixed point.** In the same-pair series the per-round increment converges to exactly 0: the state reaches a value that the next round leaves bit-identical, so the error stops growing (7.15×10^{-6} from K=128 onward) rather than accumulating.

We report the sublinear slopes (0.48–0.67) and saturation as measured properties of this implementation (libm and rounding specifics), and the $K\mathrm{u}$ envelope as the theorem's claim.

D Gated undo derivation and kernel conventions

Kernel convention. The reference gated kernel executes

$$S \;\mapsto\; \alpha\,(I - \beta k k^\top)\,S + \beta\,k v^\top, \qquad \alpha = \exp(g) > 0,$$

i.e. the gate multiplies *only* the old state; the write term $\beta k v^\top$ is not gated (verified against the released fused recurrent kernel). This convention matters: a gate multiplying the whole right-hand side would change the undo below.

Uniqueness of the same-form gated undo. Seek an undo step of the same functional form, $S \mapsto \alpha'(I - \beta'' k k^\top)S + \beta'' k \tilde{v}^\top$, whose composition with the forward step is the identity for all S. Write $s = \|k\|_2^2$. *Linear part.* On $k^\perp$ the composite acts as $\alpha'\alpha$; along k as $\alpha'\alpha(1 - \beta'' s)(1 - \beta s)$. The identity forces $\alpha'\alpha = 1$, then $(1 - \beta'' s)(1 - \beta s) = 1$, i.e.

$$\alpha' = 1/\alpha, \qquad \beta'' = \beta' = -\tfrac{\beta}{1 - \beta s}. \tag{7}$$

Affine part. With these values ($1 - \beta' s = 1/(1 - \beta s)$ and $\alpha' = 1/\alpha$), $\alpha'(I - \beta' k k^\top)\,\beta k v^\top = \alpha'\beta(1 - \beta' s)\,k v^\top = -\tfrac{\beta'}{\alpha}\,k v^\top$, so the composite's affine term is $-\tfrac{\beta'}{\alpha}\,k v^\top + \beta' k \tilde{v}^\top$, which vanishes for all v iff

$$\tilde{v} = \alpha' v = v/\alpha.$$

This proves Theorem 6(d): the same-form gated undo exists, is unique, and rescales the value vector by $1/\alpha$.

Measured (three states; `results/port_results.json`, t5). No-gate configuration $\alpha \equiv 1$ (the state-tracking setting): undo error 1.24×10^{-16} (fp64), 7.41×10^{-8} (fp32). General gated undo $(k, \tilde{v} = v/\alpha, \beta', \alpha' = 1/\alpha)$: 1.78×10^{-16} (fp64), 1.67×10^{-7} (fp32). The best *in-domain* attempt ($\alpha' = \alpha$, same v) fails at 0.89; using the same v without rescaling fails at 0.43—both separated from the correct formula by many orders of magnitude, confirming that the value rescaling is forced by the algebra, not a modeling choice.

Porting cost recap. (i) $\alpha' = 1/\alpha \geq 1$ leaves the training domain $\alpha \in (0, 1]$ (free only at $\alpha = 1$—precisely the no-forget-gate state-tracking configuration); (ii) β' carries the sign/spectrum price of Theorem 6(c); (iii) $\tilde{v} = v/\alpha$ rescales a free value vector—no additional domain cost.