# LEARNING TO REMEMBER WITH MÖBIUS UPDATES

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Recurrent sequence models with matrix-valued hidden states provide fixed-size associative memory, but their update rules affect what information is kept. We study a family of linear-fractional memory updates that keep the state bounded in spectral norm, including affine updates and nonlinear updates induced by matrix Riccati dynamics. We give sufficient conditions for boundedness, analyze memory writing and retention, and derive a chunkwise form exactly equivalent to the sequential recurrence. Experiments on synthetic associative recall and state-tracking tasks examine accumulation, overwriting, and deletion, as well as generalization to longer sequences and increased memory load. Results show improvements over recurrent baselines in selected settings, with the relative performance of affine and nonlinear variants depending on the task and parameterization.

## 1 INTRODUCTION

A recurrent matrix memory stores associations in a fixed number of coordinates and rewrites those same coordinates at every token (Katharopoulos et al., 2020; Schlag et al., 2021). What such a memory can compute is a property of the whole family of writes that act in this one shared representation. A single nonlinear update can become affine after a change of coordinates, yet every later write must use the same coordinates. The delta rule already corrects each write with stored content and stays affine at fixed controls (Yang et al., 2024b), and gated and signed transitions enlarge this affine family (Hatamizadeh et al., 2026; Grazzi et al., 2025; Siems et al., 2026). We ask which interactions among stored contents a bounded nonlinear write can add, whether such a write still admits exact parallel training, and whether trained networks use the added interactions.

We study matrix memories whose writes are Möbius transformations of a spectral ball, while queries read the stored content linearly (Figure 1). At fixed controls, every write equals its affine tangent at empty memory plus a rank-one remainder, the product of two reads of stored content. In bases aligned with the written value and key, the remainder performs three operations. It changes how the written association accepts new content, rescales the cross-associations that share the written value or key, and adds a scaled product of these cross-associations to the content outside both write directions. The last operation acts even when the written association is empty.

The family of such writes has two structural properties that single formulas do not reveal. Writes along one fixed value and key pair become translations in explicit coordinates. A finite family of writes along different directions, by contrast, has no common continuous injective encoding of the same dimension in which every write is affine (Proposition 1). The same fractional structure admits an exact chunkwise algorithm with one unknown per token that returns every causal prefix read of the sequential recurrence (Section 5).

In matched Chess World Model training (Walker & Lyons, 2026), the Möbius write improves Full75 state accuracy over native Gated DeltaNet-2 by a mean of 1.55 and 1.53 percentage points on held-out May and June games, with a positive gain in every run. Against its matched tangent, it improves Board64 in all three paired runs on May, June and random legal games, while Full75 differences are smaller and mixed. Frozen networks show where the nonlinear computation acts. The remainder component outside both write directions is the product of two lateral stored associations, and removing its accumulated history from the final-layer read lowers Board64 from 87.8% to 72.9%. Removing the rest of the remainder leaves 86.7%, while both removals break the full-move counter, whose accuracy falls from 98.6% to 39.9% and 53.4%. Replacing the write in a trained curved network by its matched tangent lowers June Full75 by 60.9 points, yet a tangent trained from the start comes within 0.6 points of Curved.



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.

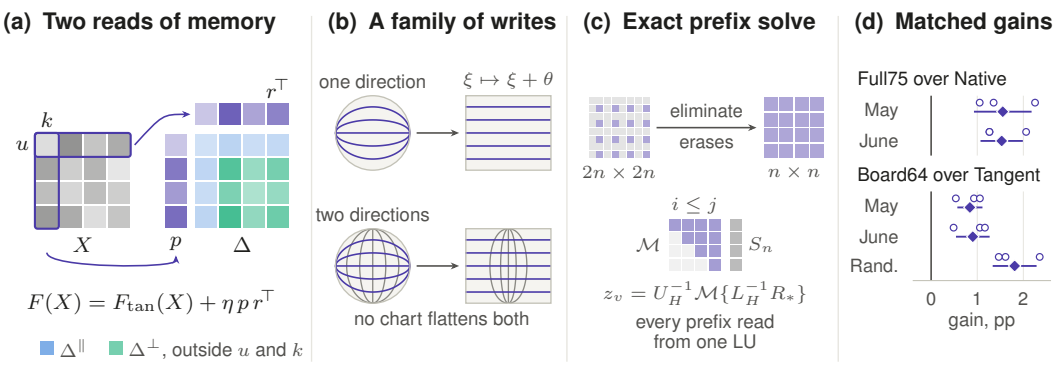


Figure 1: **Overview.** (a) A Möbius write adds a scaled product of two reads of stored memory to its matched affine tangent. (b,c) The family of such writes resists common affine coordinates yet admits an exact chunkwise solve. (d) Per-run gains of Curved in matched training.

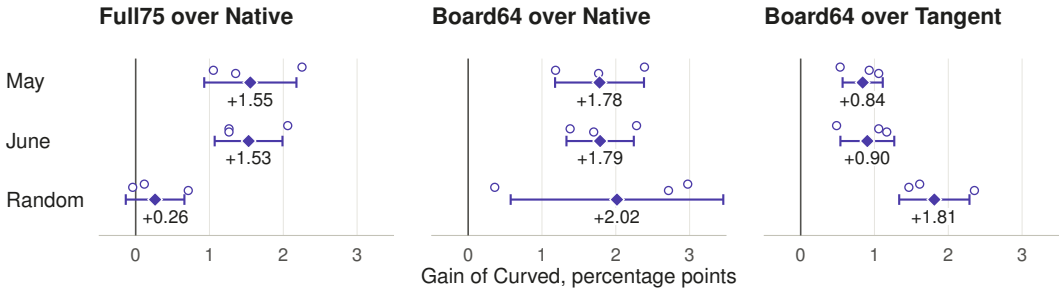


Figure 2: **Gains of Curved after 128M training moves.** Paired differences in percentage points. Circles are the three runs, and diamonds show the mean with one sample SD.

## 2 A MATCHED STUDY OF MEMORY WRITING

Chess World Model (CWM) predicts 75 state labels after every move of a game from the move sequence alone (Walker & Lyons, 2026). The model must maintain the 64 square labels, side to move, castling rights, the en passant square and two move counters through complete games. *Board64* counts a state as correct when all 64 square labels are correct, and *Full75* when all 75 labels are. Joint accuracy rewards a consistent state, the property a memory has to maintain.

We train native Gated DeltaNet-2 (Hatamizadeh et al., 2026), the Möbius memory of Section 3 and its matched affine tangent inside the same frontend and task model, and call them Native, Curved and Tangent. Curved and Tangent have 7,733,471 parameters each, including one shared curvature scalar, and Native has one parameter fewer. Each model is trained three times on 128M moves from one corpus of 1M March games, with paired initializations and batch order. Tangent keeps the curvature parameterization and replaces each write by its value and Jacobian at empty memory, so the contrast $C - T$ isolates the availability of the nonlinear remainder. May and random legal games are held-out cohorts, and June games form a confirmation cohort fixed before evaluation (Appendix D.1).

Curved improves Full75 over Native by $1.55 \pm 0.62$ percentage points (pp) on May and $1.53 \pm 0.46$ pp on June, positive in all three pairs on both cohorts (Table 1, Figure 2). Against Tangent, Curved improves Board64 by $0.84 \pm 0.27$, $0.90 \pm 0.36$ and $1.81 \pm 0.48$ pp on May, June and Random. Each of the three training pairs improves Board64 on every cohort. The corresponding Full75 differences are $0.57 \pm 0.74$, $0.59 \pm 0.84$ and $-0.14 \pm 0.66$ pp, with mixed signs across pairs.

The Random cohort shows why the two joint metrics diverge. There Curved raises the mean accuracy of every field group and raises Board64 by 1.81 pp, while Full75 changes by $-0.14$ pp. Among states with a correct board, 42.0% of Curved states and 34.0% of Tangent states contain an auxiliary

Table 1: State accuracy after 128M training moves, in percent. Mean and sample SD over three matched runs.

| | Full75 | | | Board64 | | |
|---|---|---|---|---|---|---|
| Cohort | Native | Tangent | Curved | Native | Tangent | Curved |
| May | $81.94 \pm 0.41$ | $82.92 \pm 0.94$ | $83.50 \pm 0.24$ | $84.38 \pm 0.06$ | $85.32 \pm 0.77$ | $86.16 \pm 0.55$ |
| June | $82.39 \pm 0.35$ | $83.34 \pm 0.95$ | $83.92 \pm 0.14$ | $84.78 \pm 0.07$ | $85.67 \pm 0.79$ | $86.57 \pm 0.47$ |
| Random | $9.33 \pm 0.11$ | $9.73 \pm 0.33$ | $9.59 \pm 0.39$ | $14.54 \pm 0.81$ | $14.75 \pm 0.87$ | $16.56 \pm 0.79$ |

error. Field accuracy, board accuracy and joint accuracy therefore measure different properties of the tracked state, and the gain from the nonlinear write concentrates in the joint board configuration. Section 6 asks which part of the write carries board and counter computation. Appendix E reports synthetic recall, language modeling and pretrained hybrid adaptation, where the ranking of update families depends on the task and the trainable subspace.

## 3 What a Möbius write changes

A fixed-size associative memory stores new and old information in the same coordinates. We describe what one write does to an occupied state. Let $S \in \mathbb{R}^{V \times K}$ with Euclidean vector norms and the spectral matrix norm. Each token computes its controls causally from the layer input, and all operation comparisons below hold these controls fixed. A diagonal decay $0 \preceq D \preceq I$, an erase gate $b \in [0,2]^K$, a key $k$ with $\|k\| \le 1$, a value $u$ and a query $q$ give

$$e = \sqrt{b} \odot k, \quad A = D(I - ee^\top), \quad X = S_{t-1}A, \qquad S_t = F_{\kappa,u,k}(X), \quad y_t = S_t q. \tag{1}$$

The contraction satisfies $\|A\| \le 1$, including signed erasure. We keep the GDN2 frontend with its gates, short convolutions and output normalization (Hatamizadeh et al., 2026), and use the symmetric eraser above. Curvature $\kappa = a^2$ is one learned scalar shared across heads and layers, initialized at $a = 1/8$. With $c = \sqrt{1 + \kappa\|u\|^2\|k\|^2}$, the write is

$$F_{\kappa,u,k}(X) = (PX + uk^\top)(\kappa k u^\top X + Q)^{-1}, \qquad P = I + \frac{\kappa\|k\|^2 uu^\top}{c+1}, \quad Q = I + \frac{\kappa\|u\|^2 kk^\top}{c+1}. \tag{2}$$

At $\kappa = 0$ this is the additive write $X + uk^\top$. For $\kappa > 0$ it maps the ball $\mathcal{B}_\kappa = \{S : \|S\| < 1/\sqrt{\kappa}\}$ onto itself, as a classical automorphism of the matrix ball (Harris, 1974; 1992), and infinitesimal writes follow a matrix Riccati equation (Appendix B.1). Kalman Delta Networks scan Möbius updates of an uncertainty state and keep the content recurrence affine at fixed gains (Bui et al., 2026). Here the fractional map acts on the queried content itself.

**Tangent and remainder.** The matched tangent keeps the same curvature and controls. With

$$\beta = \frac{\kappa\|k\|^2}{c(c+1)}, \quad \alpha = \frac{\kappa\|u\|^2}{c(c+1)}, \quad w = \frac{1}{c}, \quad g = \frac{\kappa}{c}, \quad L = I - \beta uu^\top, \quad R = I - \alpha kk^\top, \tag{3}$$

the write splits exactly as

$$F_\kappa(X) = \underbrace{LXR + wuk^\top}_{F_{\tan}(X) = F_\kappa(0) + DF_\kappa(0)[X]} + \Delta, \qquad \Delta = \eta\, \underbrace{(LXk)}_{p}\, \underbrace{(u^\top XR)}_{r^\top}, \quad \eta = -\frac{g}{1 + g\, u^\top Xk}. \tag{4}$$

The tangent matches the curved write in value and first derivative at empty memory. The remainder multiplies two reads of stored content. The column $p$ reads the values stored at the current key, and the row $r^\top$ reads the keys associated with the current value. At a query $q$, $\Delta q = \eta p\,(r^\top q)$, so existing content selects which keys the remainder reaches, including keys orthogonal to $k$.

**Three operations inside one write.** For $\kappa > 0$ and nonzero $u, k$, express the normalized state $M = \sqrt{\kappa}X$ in orthonormal value and key bases whose first vectors are $u/\|u\|$ and $k/\|k\|$ (Figure 3).

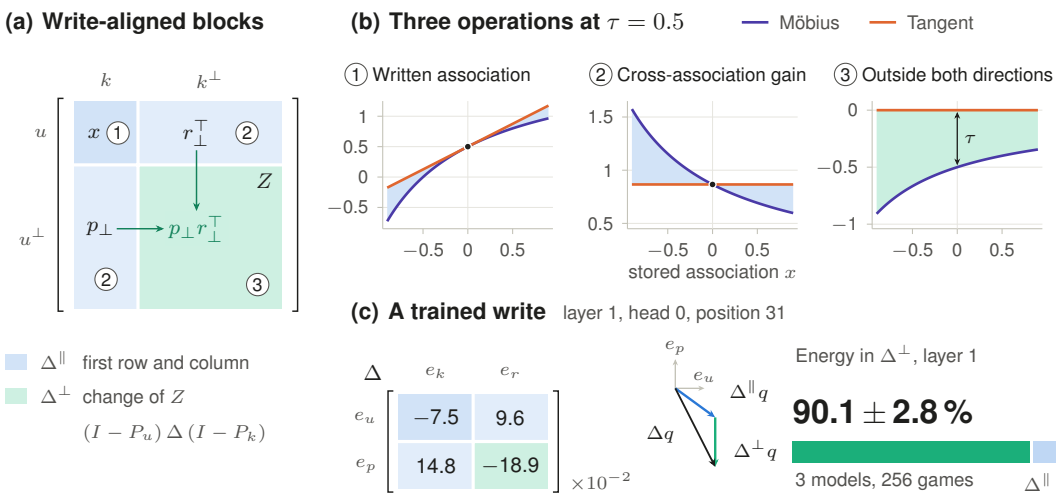


Figure 3: **Three operations inside one write.** (a) Blocks of the state in bases aligned with the write. (b) Curved and tangent maps at $\tau = 0.5$ as functions of the stored association $x$. (c) A layer-1 correction from a trained checkpoint in its two active planes, and the share of layer-1 correction energy in $\Delta^{\perp}$.

Let $\tau = \sqrt{\kappa}\|u\|\|k\|/c$ and $h = \sqrt{1-\tau^2} = 1/c$. The curved and tangent maps act blockwise as

$$M = \begin{pmatrix} x & r_\perp^\top \\ p_\perp & Z \end{pmatrix}, \quad \widehat{F}_\tau(M) = \begin{pmatrix} \dfrac{x+\tau}{1+\tau x} & \dfrac{h\, r_\perp^\top}{1+\tau x} \\ \dfrac{h\, p_\perp}{1+\tau x} & Z - \dfrac{\tau\, p_\perp r_\perp^\top}{1+\tau x} \end{pmatrix}, \quad \widehat{T}_\tau(M) = \begin{pmatrix} \tau + h^2 x & h\, r_\perp^\top \\ h\, p_\perp & Z \end{pmatrix}. \tag{5}$$

The scalar $x$ is the stored association between the written value and key, and its fractional update changes how an occupied coordinate accepts a new write. The blocks $p_\perp$ and $r_\perp$ are cross-associations that share the written key or value, and the curved write scales them by $h/(1+\tau x)$, a gain set by the stored association. The block $Z$ holds all content outside both write directions. The curved write adds a scaled $p_\perp r_\perp^\top$ to it, so two existing cross-associations change a third association elsewhere in memory. At $x = 0$ the two maps differ only in $Z$, by $-\tau p_\perp r_\perp^\top$, and this interaction acts even when the written association is empty.

**The measurable component.** Projecting the remainder away from both write directions isolates the third operation. With $P_v = vv^\top/\|v\|^2$ for $v \neq 0$ and $P_0 = 0$, the identities $(I - P_u)L = I - P_u$ and $R(I - P_k) = I - P_k$ give

$$\Delta^{\perp} = (I - P_u)\,\Delta\,(I - P_k) = -\frac{g}{1 + g\, u^\top X k}\,\big[(I - P_u)Xk\big]\big[u^\top X (I - P_k)\big]. \tag{6}$$

Its factors are the stored values at key $k$ that lie outside $u$ and the stored key pattern of $u$ that lies outside $k$. In (5), $\sqrt{\kappa}\,\Delta^{\perp}$ is exactly the change of $Z$, and $\Delta^{\parallel} = \Delta - \Delta^{\perp}$ collects the changes of $x$, $p_\perp$ and $r_\perp$. The three operations thus enter experiments as two measurable groups, and Section 6 follows both through trained networks.

## 4 Geometry of a Family of Writes

**One direction pair flattens.** At fixed $\kappa > 0$, pure writes with $A = I$ along one oriented value and key pair admit explicit coordinates on the whole ball in which a write of signed strength $\tau \in (-1, 1)$ translates one coordinate, $\xi \mapsto \xi + \operatorname{atanh}\tau$, and fixes all others (Appendix B.2). A single nonlinear formula therefore does not determine what a memory computes.

**A family of directions does not.** Let $\mathbb{B} = \{X \in \mathbb{R}^{V\times K} : \|X\| < 1\}$ be the normalized ball, $\mathbb{B}_r$ its closed subball of radius $r$, and $s \cdot X$ the action of a pure write $s$. A coordinate write uses a value and a key along coordinate axes.

**Proposition 1.** *Let $V \geq 2$, $K \geq 1$ and $n = VK$. For every $0 < r < 1$ there is a set $\mathcal{S}_r$ of $4VK$ coordinate writes, fixed independently of any encoding, such that every continuous injective $F : \mathbb{B} \to \mathbb{R}^n$ satisfies*

$$\delta_{F,r} = \inf_{\{\Psi_s\}} \max_{s\in\mathcal{S}_r} \sup_{X\in\mathbb{B}_r} \|\Psi_s(F(X)) - F(s \cdot X)\|_2 > 0,$$

*where each write $s$ receives its own arbitrary affine map $\Psi_s(z) = L_s z + b_s$ of $\mathbb{R}^n$.*

The gap is positive for every encoding and depends on it. The proof proceeds in two steps (Appendix B.6). A zero gap yields exact affine maps for the finite writes, and incommensurate step sizes extend them to every small coordinate write, giving a local affine action of the whole write family. The commutators of coordinate writes generate the semisimple Lie algebra $\mathfrak{so}(V, K)$ of the ball. Complete reducibility then produces a common fixed point, and the isotropy at that point contradicts injectivity. Preceding every write by a slight strict contraction preserves a positive gap, so the result also covers the decaying recurrence.

A larger homogeneous representation removes the obstruction. Writing $S = PQ^{-1}$ with a frame $(P, Q)$ of $(V + K)K$ coordinates, every write acts linearly on the frame, and the read requires the fractional map $PQ^{-1}$. Nonlinearity can thus move between state size, update and readout. The Möbius memory keeps $VK$ coordinates and a linear read, and places the nonlinearity in the update. Proposition 1 characterizes the operations available to a controller, and Section 6 measures which of them trained networks use.

**Further structure.** Three suitable pure writes at one key can compose exactly to a rotation of every value column, $S \mapsto \mathcal{R}S$, a Thomas–Wigner rotation (Başkal & Kim, 2002; Ungar, 2008); affine transitions can also realize rotations (Siems et al., 2026), and Proposition 1 concerns the whole family (Appendix B.6). The cross-association gain alone already separates normalized reads. An eight-state construction with $p_\perp = 0$ leaves $Z$ unchanged and still has a positive gap from every affine state map after RMS normalization (Proposition 5). With strict decay $\|A_t\| \leq \rho < 1$ and bounded controls, the curved state stays strictly inside its ball with a sharp uniform radius, and two trajectories with equal controls contract in the invariant distance of the ball by $\prod_j \|A_j\|$ (Propositions 2 and 3).

## 5 Exact chunkwise computation

Parallel training requires all reads of a chunk of $n$ tokens at once. Each token applies an erasing contraction followed by a rank-one write. Treating both as unknowns gives a system of size $2n$. Eliminating the erase unknowns with one triangular solve leaves one write unknown per token. Appendix C builds, from direct interval products of the decays, a matrix $H \in \mathbb{R}^{n\times n}$ and right-hand sides $R_* \in \mathbb{R}^{n\times(n+K)}$ whose first $n$ columns belong to the $n$ queries and whose last $K$ columns belong to the final state. The chunk is then solved as

$$H = L_H U_H, \qquad z_v = U_H^{-1}\,\mathcal{M}\{L_H^{-1} R_*\}, \tag{7}$$

where $L_H$ is unit lower triangular and $\mathcal{M}$ multiplies entry $(i, j)$ of a query column by $1\{i \leq j\}$, leaving the final-state columns unchanged. The reads $y_1, \ldots, y_n$ and the final state $S_n$ follow from $z_v$ by one matrix product (Appendix C).

The placement of the mask makes every prefix exact. The read at token $j$ must depend only on writes $i \leq j$, which is the leading $j \times j$ system of $H$. The leading block of $H$ factors as the leading blocks of $L_H$ and $U_H$, and the lower solve for rows $1, \ldots, j$ ignores later rows. Zeroing rows $i > j$ between the two solves then makes the upper solve return exactly the $j$-token solution, for all $j$ in one pass. Without the mask, the dense coupling in $H$ would send contributions from later writes into earlier reads. The leading minors are products of the sequential write pivots, $\det H_{1:r,1:r} = \prod_{i\leq r}(1 + g_i u_i^\top X_i k_i) > 0$, so unpivoted LU exists and the prefix structure is the one it preserves (Proposition 6).

Table 2: Agreement of the chunk solve with the sequential recurrence, as relative maximum error.

| Quantity | FP64 | FP32 |
|---|---|---|
| All reads | $4.3 \times 10^{-16}$ | $2.4 \times 10^{-7}$ |
| Final state | $8.5 \times 10^{-16}$ | $5.7 \times 10^{-7}$ |
| Input gradients | $2.5 \times 10^{-15}$ | $1.5 \times 10^{-6}$ |
| Curvature gradient | $2.9 \times 10^{-14}$ | $4.9 \times 10^{-6}$ |

Table 3: Forward and backward pass on one batch at $n = 32$, with memory above the pre-forward baseline.

| Model | ms | MiB | Solver |
|---|---|---|---|
| Native | 14.40 | 141.6 | FLA kernel |
| Tangent | 51.44 | 861.9 | LU |
| Tangent | 43.46 | 855.7 | Triangular |
| Curved | 51.95 | 727.9 | LU |

Newton and quasi-Newton methods parallelize general nonlinear recurrences by iterating the whole sequence to a fixed point (Lim et al., 2024; Gonzalez et al., 2024; Danieli et al., 2026). The fractional structure here returns every prefix of the specified recurrence through one finite factorization, with no outer convergence loop. Trellis computes a nonlinear content memory in chunks by evaluating gradients at the chunk entrance (Karami et al., 2025), whereas the solve above reproduces the sequential recurrence itself. At $\kappa = 0$, $H = I$. For the matched tangent, $g = 0$ makes $H$ unit lower triangular, so two triangular solves suffice.

Table 2 compares the chunk solve with the direct recurrence on saved controls of a trained layer at chunk lengths 32 and 64, including signed erasure, exact zero decays and the affine boundary. Reads, final states, gradients for the initial state and all five per-token inputs, and the curvature gradient agree to roundoff in both precisions. With bounded controls and a uniform interior margin of the state, $\mathrm{cond}_2(H) = O(n^4)$ (Appendix C.1). Table 3 measures a complete forward and backward pass at $n = 32$ on an RTX PRO 6000. The Curved model takes $3.61\times$ the native time and $5.14\times$ its incremental memory. Using the triangular structure of the tangent already saves $1.18\times$, and a structured dense solve for $H$ is the corresponding target for the curved write (Appendix D.3).

## 6 From writes to learned computation

Proposition 1 describes the operations a controller can use. Frozen checkpoints show which of them carry task computation. Two diagnostics use the actual controls of each network on the first 256 June games, and complete-network swaps of the write then test dependence on all 10,000 June games (Appendix D.2).

**Feedback persists across writes.** Beside each curved layer we run a tangent trajectory $\widetilde{S}_t$ with the same controls and the same initial state. Its difference $E_t = S_t - \widetilde{S}_t$ obeys the exact recursion

$$E_t = L_t E_{t-1} A_t R_t + \Delta_t, \tag{8}$$

so the current read $E_t q_t$ contains the new remainder and the remainders of earlier writes transported by the tangent. Figure 4 shows two learned regimes. In layer 0, the one-step read change has relative RMS 1.09%, while the persistent difference reaches 6.40%, and transported read energy exceeds new-remainder energy in every run. A small change per write thus accumulates into a lasting difference in memory. In layer 1, the one-step change is 35.07% and the persistent difference 44.57%, and the state-energy cross term between transported and new remainders is negative in every run, so new feedback partly cancels its history. After the actual gate, normalization and output projection, the persistent differences remain 5.77% and 44.50%. A tangent-trained layer 1 responds far less to the restored nonlinear remainder, with a one-step read change of 3.95%.

**The two remainder components carry different fields.** Following (6), we split each remainder into $\Delta_t^{\perp}$ and $\Delta_t^{\parallel}$, which are Frobenius orthogonal. In layer 1, $\Delta^{\perp}$ carries $90.1 \pm 2.8\%$ of the instantaneous correction energy. Its history $E_t^{\perp} = L_t E_{t-1}^{\perp} A_t R_t + \Delta_t^{\perp}$ uses the transport of (8), and $E_t^{\parallel} = E_t - E_t^{\perp}$. Each intervention removes one history from one layer's read, $S_t q_t \mapsto (S_t - E_t^{\perp}) q_t$ or $(S_t - E_t^{\parallel}) q_t$, and runs the actual gate, normalization, projection, residual path, MLP and the rest of the network.

In the final layer the two components separate (Figure 5c, Table 4). Removing $E^{\perp}$ lowers Board64 from 87.81% to 72.93%, while removing $E^{\parallel}$ leaves 86.68%, an ordering that holds in all three runs.

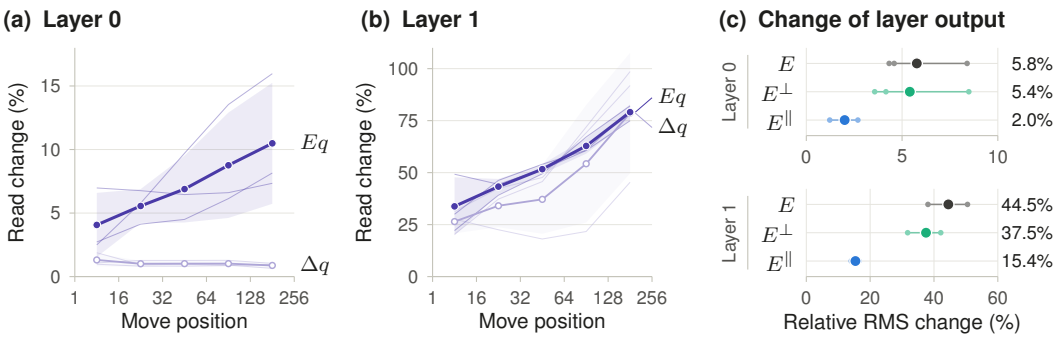


Figure 4: **Nonlinear feedback persists into the learned readout.** (a,b) One-step and persistent read changes of the curved models by move position. (c) Change of the layer output after removing each remainder history. Thin lines and small dots are runs; bands and error bars show one sample SD across three runs.

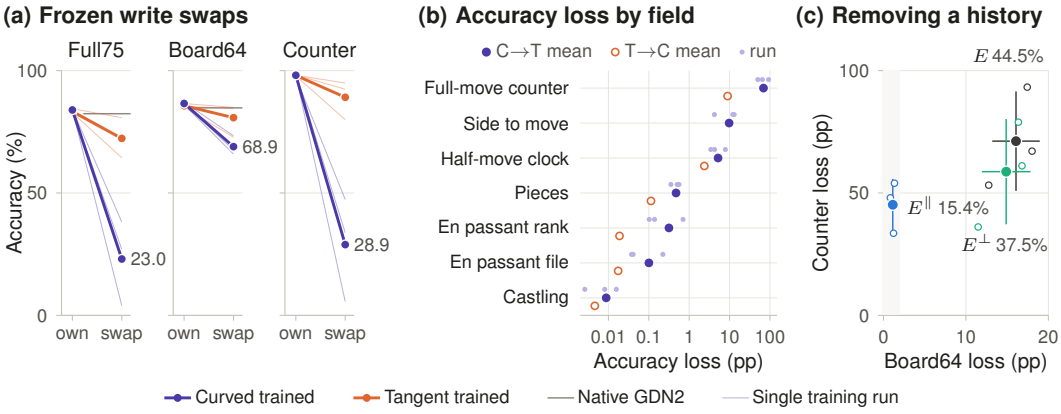


Figure 5: **Write interventions reveal field-specific dependence.** (a,b) Swapping the write of frozen networks on 10,000 June games. (c) Removing each remainder history from the final-layer read on 256 June games. Counter denotes the complete full-move counter. In (c), points and error bars show the mean and one sample SD across three runs; labels give relative RMS layer-output changes.

The full-move counter is encoded as two base-256 digits, and the high digit is correct for every state in every branch, so counter accuracy equals the accuracy of the low digit. It falls from 98.56% to 39.85% without $E^{\perp}$ and to 53.37% without $E^{\parallel}$. The interaction outside both write directions therefore supports board tracking, and both components support counting.

The two removals change the layer output by different amounts, 37.5% and 15.4% relative RMS. Per percent of output change, removing $E^{\perp}$ costs 0.40 pp of Board64 and removing $E^{\parallel}$ costs 0.07 pp, while the counter loses 1.53 and 2.95 pp. These ratios are computed within each run before averaging. Both orderings hold in every run: under these interventions, per unit of output change the board is more sensitive to $E^{\perp}$ and the counter to $E^{\parallel}$. Removing $E^{\perp}$ also lowers accuracy both on squares whose piece changes at the current move, by 1.88 pp, and on the far more numerous unchanged squares, by 0.32 pp, so the board computation involves updating and retaining pieces alike (Appendix Table 4).

**Changing the write changes the learned strategy.** A complete-network swap sets $g = 0$ in every write of a curved-trained network, keeping all weights and tangent coefficients and recomputing later controls, and the reverse swap restores $g$ in a tangent-trained network. Removing the nonlinear remainder lowers June Full75 by $60.90 \pm 17.43$ pp and the complete counter from 98.10% to 28.90% (Figure 5). Side-to-move accuracy falls from 99.99% to 90.24% and piece-label accuracy from 99.73% to 99.25%. The identity $\Delta\text{Full75} = \Delta\text{Board64} - \Delta P(\text{board correct, auxiliary wrong})$

assigns 43.27 of the 60.90 points to states with a correct board and an incorrect auxiliary label. Restoring the nonlinear remainder to tangent-trained weights lowers Full75 by $11.03 \pm 7.35$ pp and the counter from 98.04% to 89.10%.

Independent training lets the tangent learn another computation. Its June Full75 is 83.34%, against 83.92% for Curved and 82.39% for Native. A curved-trained network depends heavily on the remainder, and a tangent trained from the start recovers most of that accuracy through different controls. Learned use of an operation and the gain from offering it before training are therefore distinct measurements.

## 7 Related work

Delta-rule memories correct each write with stored content (Yang et al., 2024b; 2025), and GDN2, Kimi Delta Attention and Erase-then-Delta Attention refine gates, erasure and addressing (Hatamizadeh et al., 2026; Kimi Team et al., 2025; Li et al., 2026). At fixed controls these updates remain affine in the state. Signed eigenvalues, Householder products and the rotations of CKDA enlarge the affine family and its state-tracking power (Merrill et al., 2024; Grazzi et al., 2025; Siems et al., 2025; 2026). Proposition 1 concerns a different axis, the common representation of a nonlinear write family, and holds against arbitrary affine comparators. Kalman Delta Networks scan Möbius updates of an uncertainty state (Bui et al., 2026), whereas the fractional map here acts on the queried content. Test-Time Training, Titans and Trellis learn nonlinear content updates (Sun et al., 2024; Behrouz et al., 2025; Karami et al., 2025), and Section 5 relates the exact solve to chunked and Newton-type parallel evaluation. Matrix-ball automorphisms, their invariant metrics and Riccati flows are classical (Harris, 1974; 1992; Ostrovskii et al., 2009; Arov & Dym, 2008). This work connects them to associative content, a measurable remainder, shared-coordinate expressivity and exact chunk evaluation. Appendix A gives a broader account.

## 8 Discussion

**Where should content interaction live?** At fixed state size and matched compute, the remainder built from two reads can be compared with a wider or deeper controller or with several affine updates per token. Matching training and inference costs separately would show when a richer write supplies useful computation more cheaply than a larger controller.

**Which operations does a task require?** Tasks that vary time counting, coordination between stored associations, overwriting and deletion independently can test predictions made before training about which remainder component affects which output. A direct test is to match the two removals in output magnitude and ask whether the board remains more sensitive to $E^{\perp}$ and the counter to $E^{\parallel}$.

**Which update family comes next?** Direct content reads, a characterized operation family and an exact structured solver can be required together. Joint measurements of conditioning, finite-precision error, capacity, retention and cost would then decide which interactions stay useful at a practical budget, and the dense LU of the present prototype sets a concrete kernel target.

## REPRODUCIBILITY STATEMENT

Appendices B and C provide the assumptions, proofs and full chunk algorithm. Appendices D and E describe the experimental protocols. The anonymous supplementary code includes the implementations, configurations, retained numerical evidence, and commands for checking the mathematics and reconstructing reported summaries. Its README distinguishes replay of retained evidence from new training runs and documents the external data and checkpoint requirements; original training checkpoints are not included.

## AI USE STATEMENT

The central view of this work, that the computation available to a memory write is a property of the operation family acting in one shared state representation, was formed and led by the authors. The authors developed it, together with the hypotheses that the Möbius remainder carries task-relevant interactions and that its fractional structure admits an exact chunkwise solve, through discussions with GPT-5.6 Pro. The mathematical claims and their proofs, including Proposition 1, the normalized-read gap of Proposition 5 and the conditioning bound of Proposition 7, were formulated and written with assistance from GPT-6 Astra (Max), and the authors checked, corrected and edited every proof. The code for the recurrences, the exact chunk solver, data preparation, training, evaluation and the mechanism diagnostics was written with assistance from GPT-5.6 Sol (Ultra) and GPT-6 Astra (Xhigh), which also gave feedback on the design of the matched experiments. Data generation and the interpretation of results were led by the authors, and GPT-5.6 Sol (Ultra) wrote the programs that analyze the retained outputs. The authors verified the chunk solver against the direct recurrence (Table 2), fixed every evaluation cohort before its evaluation, and checked each reported number against the retained result files. The authors wrote the main body of the paper. Claude Opus 5.5 (Xhigh) and GPT-6 Astra assisted with polishing and restructuring the text, typesetting, drawing the figures from retained results, identifying and formatting references, and computing descriptive summaries of retained outputs, such as the complete counter accuracy and the per-unit sensitivities in Section 6. The authors recomputed or checked these summaries against the retained outputs. We have not used generative AI tools to generate data sets, to interpret results or for translation; qualitative or thematic data analysis is not applicable to this work. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

## REFERENCES


Simran Arora, Sabri Eyuboglu, Aman Timalsina, Isys Johnson, Michael Poli, James Y Zou, Atri Rudra, and Christopher Ré. Zoology: Measuring and improving recall in efficient language models. In *International conference on learning representations*, volume 2024, pp. 15664–15730, 2024.

Damir Z Arov and Harry Dym. *J-contractive matrix valued functions and related topics*. Number 116 in Encyclopedia of Mathematics and its Applications. Cambridge University Press, 2008.

Jimmy Ba, Geoffrey E Hinton, Volodymyr Mnih, Joel Z Leibo, and Catalin Ionescu. Using fast weights to attend to the recent past. *Advances in neural information processing systems*, 29, 2016.

S Başkal and YS Kim. Wigner rotations in laser cavities. *Physical Review E*, 66(2):026604, 2002.

Maximilian Beck, Korbinian Pöppel, Markus Spanring, Andreas Auer, Oleksandra Prudnikova, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. xLSTM: Extended long short-term memory. *Advances in Neural Information Processing Systems*, 37:107547–107603, 2024.

Ali Behrouz, Peilin Zhong, and Vahab Mirrokni. Titans: Learning to memorize at test time. *Advances in Neural Information Processing Systems*, 38:113506–113543, 2025.

Julian Boesch and Andrew Wee. Anatomy of associative recall in fixed-state recurrences: A matched-state decomposition, an interference wall, and a curriculum that breaks it. *arXiv preprint arXiv:2609.16183*, 2026.

Ngoc Bui, Tinglin Huang, and Rex Ying. Kalman Delta Networks: Uncertainty-aware associative memory. *arXiv preprint arXiv:2609.07816*, 2026.

Yingfa Chen, Xinrong Zhang, Shengding Hu, Xu Han, Zhiyuan Liu, and Maosong Sun. Stuffed Mamba: Oversized states lead to the inability to forget. *arXiv preprint arXiv:2410.07145*, 2024.

Kyunghyun Cho, Bart Van Merriënboer, Çağlar Gulçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1724–1734, 2014.

Brian Conrad and Aaron Landesman. Math 210C. Compact Lie groups. Stanford University lecture notes, 2018. URL https://math.stanford.edu/~conrad/210CPage/handouts/lie_groups_notes.pdf.

Carl C Cowen and Barbara D MacCluer. Linear fractional maps of the ball and their composition operators. *Acta Sci. Math.(Szeged)*, 66(1-2):351–376, 2000.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G Carbonell, Quoc Le, and Ruslan Salakhutdinov. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 2978–2988, 2019.

Federico Danieli, Pau Rodríguez, Miguel Sarabia, Xavier Suau, and Luca Zappella. ParaRNN: Unlocking parallel training of nonlinear RNNs for large language models. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=mX8b64iUaa.

Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060*, 2024.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. *Advances in neural information processing systems*, 35:16344–16359, 2022.

Matthew Dowling, Hyungju Jeon, Cristina Savin, and Il Memming Park. Memory by design: Probabilistic sequence layers. *arXiv preprint arXiv:2605.31163*, 2026.

Pavel Etingof. Lie groups and Lie algebras. *arXiv preprint arXiv:2201.09397*, 2022.

Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic neural networks. *Advances in neural information processing systems*, 31, 2018.

Yun Gao, Sui-Chung Ng, and Aeryeong Seo. Generalizations of linear fractional maps for classical symmetric domains and related fixed point theorems for generalized balls. *Annali di Matematica Pura ed Applicata (1923-)*, 200(1):81–99, 2021.

Ayoub Ghriss and Sourav Chakraborty. Kernelized linear attention: Breaking the capacity wall with symmetric cones. *arXiv preprint arXiv:2607.17419*, 2026.

Xavier Gonzalez, Andrew Warrington, Jimmy T. H. Smith, and Scott W. Linderman. Towards scalable and stable parallelization of nonlinear RNNs. In *Advances in Neural Information Processing Systems*, volume 37, pp. 5817–5849, 2024. doi: 10.52202/079017-0189. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/0b2b199fdd52089b31d3a0120e400b2a-Abstract-Conference.html.

Riccardo Grazzi, Julien Siems, Arber Zela, Jörg Franke, Frank Hutter, et al. Unlocking state-tracking in linear RNNs through negative eigenvalues. In *International Conference on Learning Representations*, volume 2025, pp. 36565–36597, 2025.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021.

Dhruman Gupta, Aritra Das, and Debayan Gupta. The query knows what to forget: A second erase direction for linear attention. *arXiv preprint arXiv:2608.13668*, 2026.

Bernard Hanzon, Martine Olivi, and Ralf LM Peeters. Balanced realizations of discrete-time stable all-pass systems and the tangential Schur algorithm. *Linear Algebra and its Applications*, 418 (2-3):793–820, 2006.

Lawrence A Harris. Bounded symmetric homogeneous domains in infinite dimensional spaces. In *Proceedings on Infinite Dimensional Holomorphy: University of Kentucky 1973*, volume 364 of *Lecture Notes in Mathematics*, pp. 13–40. Springer, 1974.

Lawrence A Harris. Linear fractional transformations of circular domains in operator spaces. *Indiana University mathematics journal*, 41(1):125–147, 1992.

Ali Hatamizadeh, Yejin Choi, and Jan Kautz. Gated DeltaNet-2: Decoupling erase and write in linear attention. *arXiv preprint arXiv:2605.22791*, 2026.

Allen Hatcher. *Algebraic topology*. Cambridge University Press, 2002.

Mahdi Karami, Ali Behrouz, Praneeth Kacham, and Vahab Mirrokni. Trellis: Learning to compress key-value memory in attention models. *arXiv preprint arXiv:2512.23852*, 2025.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *International conference on machine learning*, pp. 5156–5165. PMLR, 2020.

Kimi Team, Yu Zhang, Zongyu Lin, Xingcheng Yao, Jiaxi Hu, Fanqing Meng, Chengyin Liu, Xin Men, Songlin Yang, Zhiyuan Li, et al. Kimi Linear: An expressive, efficient attention architecture. *arXiv preprint arXiv:2510.26692*, 2025.

Xiao Li, Chengruidong Zhang, Hao Luo, Xi Lin, Zekun Wang, Zihan Qiu, Yunfei Mao, Langshi Chen, Man Yuan, Minmin Sun, et al. Erase-then-delta attention: Decoupling erase and write addresses in delta-rule linear attention. *arXiv preprint arXiv:2606.26560*, 2026.

Yi Heng Lim, Qi Zhu, Joshua Selfridge, and Muhammad Firmansyah Kasim. Parallelizing non-linear sequential models over the sequence length. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=E34AlVLN0v.

William Merrill, Jackson Petty, and Ashish Sabharwal. The illusion of state in state-space models. *arXiv preprint arXiv:2404.08819*, 2024.

Antonio Orvieto, Samuel L Smith, Albert Gu, Anushan Fernando, Caglar Gulcehre, Razvan Pascanu, and Soham De. Resurrecting recurrent neural networks for long sequences. In *International conference on machine learning*, pp. 26670–26698. PMLR, 2023.

MI Ostrovskii, VS Shulman, and Lyudmila Turowska. Unitarizable representations and fixed points of groups of biholomorphic transformations of operator balls. *Journal of Functional Analysis*, 257(8):2476–2496, 2009.

Vishal Pandey and Gopal Singh. Variational linear attention: Stable associative memory for long-context transformers. *arXiv preprint arXiv:2605.11196*, 2026.

Sumin Park, Seojin Kim, and Noseong Park. Q-Delta: Beyond key-value associative state evolution. *arXiv preprint arXiv:2606.08804*, 2026.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Leon Derczynski, et al. RWKV: Reinventing RNNs for the transformer era. In *Findings of the association for computational linguistics: EMNLP 2023*, pp. 14048–14077, 2023.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, et al. Hopfield networks is all you need. *arXiv preprint arXiv:2008.02217*, 2020.

RM Redheffer. On a certain linear fractional transformation. *Journal of Mathematics and Physics*, 39(1-4):269–286, 1960.

Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. Linear transformers are secretly fast weight programmers. In *International conference on machine learning*, pp. 9355–9366. PMLR, 2021.

Vaisakh Shaj, Cameron Barker, Aidan Scannell, Andras Szecsenyi, Elliot J Crowley, and Amos Storkey. Kalman linear attention: Parallel Bayesian filtering for efficient language modelling and state tracking. *arXiv preprint arXiv:2602.10743*, 2026.

Julien Siems, Timur Carstensen, Arber Zela, Frank Hutter, Massimiliano Pontil, and Riccardo Grazzi. DeltaProduct: Improving state-tracking in linear RNNs via Householder products. *Advances in Neural Information Processing Systems*, 38:153738–153782, 2025.

Julien Siems, Riccardo Grazzi, Korbinian Pöppel, Jaisidh Singh, Arber Zela, Timur Carstensen, Jenia Jitsev, Frank Hutter, Volkan Cevher, Antonio Orvieto, et al. Complex KDA: Understanding and enhancing the expressivity of Kimi Delta Attention. *arXiv preprint arXiv:2609.24797*, 2026.

Yu Sun, Xinhao Li, Karan Dalal, Jiarui Xu, Arjun Vikram, Genghan Zhang, Yann Dubois, Xinlei Chen, Xiaolong Wang, Sanmi Koyejo, et al. Learning to (learn at test time): RNNs with expressive hidden states. *arXiv preprint arXiv:2407.04620*, 2024.

Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. Retentive network: A successor to transformer for large language models. *arXiv preprint arXiv:2307.08621*, 2023.

Abraham Albert Ungar. *Analytic hyperbolic geometry and Albert Einstein's special theory of relativity*. World Scientific, 2008.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Benjamin Walker and Terry Lyons. Chess-World-Model: A 10M-game benchmark for exact state tracking from chess move sequences. *arXiv preprint arXiv:2605.30100*, 2026.

Songlin Yang and Yu Zhang. FLA: A Triton-based library for hardware-efficient implementations of linear attention mechanism, January 2024. URL https://github.com/fla-org/flash-linear-attention.

Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. Gated linear attention transformers with hardware-efficient training. In *Proceedings of ICML*, 2024a.

Songlin Yang, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. Parallelizing linear transformers with the delta rule over sequence length. In *Proceedings of NeurIPS*, 2024b.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving Mamba2 with delta rule. In *Proceedings of ICLR*, 2025.

Chenyu Zhou, Hongpei Li, Yuerou Liu, Jianghao Lin, Dongdong Ge, and Yinyu Ye. OSDN: Improving delta rule with provable online preconditioning in linear attention. *arXiv preprint arXiv:2605.13473*, 2026.

# A Related work and the role of the memory state

**Where information is stored.** Attention stores token representations for retrieval (Vaswani et al., 2017), and Transformer-XL reuses them across segments (Dai et al., 2019). Fast weights instead store associations in parameters that change within a sequence (Ba et al., 2016). Linear attention exposes a recurrent matrix state (Katharopoulos et al., 2020); its fast-weight interpretation connects additive writes to residual correction (Schlag et al., 2021). Hopfield networks relate associative retrieval to attention (Ramsauer et al., 2020), while gated recurrent units already have nonlinear state dependence (Cho et al., 2014). Our question concerns which content interactions a directly readable, bounded matrix state supports during writing.

**Recurrence and parallel evaluation.** Structured dynamics and recurrence parameterization support long sequences in S4 and LRU (Gu et al., 2021; Orvieto et al., 2023). Mamba introduces input-dependent controls, and Mamba-2 connects structured recurrence with attention computation (Gu & Dao, 2023; Dao & Gu, 2024). GLA, RetNet and RWKV combine recurrent inference with parallel training (Yang et al., 2024a; Sun et al., 2023; Peng et al., 2023); xLSTM adds gated covariance memory (Beck et al., 2024). These architectures make the composition rule computationally important. FlashAttention and FLA further demonstrate the role of memory traffic and kernels (Dao et al., 2022; Yang & Zhang, 2024). DEER casts a nonlinear recurrence as a fixed-point problem and solves it with parallel Newton iterations (Lim et al., 2024). Quasi-DEER uses quasi-Newton updates, while ELK combines Levenberg–Marquardt with Kalman smoothing to improve stability (Gonzalez et al., 2024). ParaRNN scales Newton iterations with parallel reductions and structured-Jacobian GRU/LSTM cells to billion-parameter language models (Danieli et al., 2026). These approaches support nonlinear recurrences through iterative solves. Our erase elimination instead exploits the fractional structure of one update family and obtains every prefix of the specified recurrence from one finite factorization.

**What a write can change.** DeltaNet composes rank-one residual edits efficiently (Yang et al., 2024b), and Gated DeltaNet adds decay (Yang et al., 2025). GDN2 and Erase-then-Delta Attention separate erase/write controls and addresses, respectively (Hatamizadeh et al., 2026; Li et al., 2026); Kimi Delta Attention supplies fine-grained gates (Kimi Team et al., 2025). OSDN changes preconditioning (Zhou et al., 2026), Q-Delta mixes key/query errors (Park et al., 2026), and QED adds a query-derived erase direction (Gupta et al., 2026). These methods already consult stored content. At fixed controls, our nonlinear remainder instead multiplies two content reads, exposing both cross-association gains and a correction outside the current write directions. Its affine tangent retains the same empty-state value and linear response.

**Families in shared coordinates.** State-tracking limitations depend on the transition family (Merrill et al., 2024). Signed eigenvalues and Householder products enlarge affine families (Grazzi et al., 2025; Siems et al., 2025); CKDA realizes planar rotations through signed gates and extended delta transitions (Siems et al., 2026). The relevant distinction is therefore simultaneous representation of a write family. Our obstruction permits arbitrary affine comparators after one continuous injective encoding into the same $VK$ coordinates.

**Uncertainty and nonlinear content.** Probabilistic sequence layers retain uncertainty over associations (Dowling et al., 2026). Kalman Linear Attention uses information-form filtering (Shaj et al., 2026); KDN scans Möbius uncertainty updates before evaluating a content recurrence affine at fixed gains (Bui et al., 2026). Variational Linear Attention derives regularized least-squares updates (Pandey & Singh, 2026). Our fractional denominator acts on the directly queried content matrix. Test-Time Training and Titans learn memory within the sequence (Sun et al., 2024; Behrouz et al., 2025). Trellis also updates nonlinear content, but its chunk form uses gradients at the chunk entrance (Karami et al., 2025). Our construction preserves the specified sequential recurrence and every prefix read.

**Capacity and task structure.** Associative recall, forgetting failures, matched architectural decompositions and feature geometry expose different capacity constraints (Arora et al., 2024; Chen et al., 2024; Boesch & Wee, 2026; Ghriss & Chakraborty, 2026). CWM supplies exact state labels (Walker & Lyons, 2026). Together these perspectives motivate separating memory load, history length and

update operations, then combining endpoint accuracy with layer and field interventions to identify learned computation.

**Fractional geometry and realizations.** Matrix-ball transformations and invariant metrics are classical (Harris, 1974; 1992; Ostrovskii et al., 2009), as are fractional self-maps of vector balls and symmetric domains (Cowen & MacCluer, 2000; Gao et al., 2021). $J$-contractive functions connect this geometry to passive systems (Arov & Dym, 2008); hyperbolic networks use Möbius neural operations (Ganea et al., 2018). Our closed-write construction specializes Thomas–Wigner rotation (Başkal & Kim, 2002; Ungar, 2008). Fractional scattering composition and Schur realizations also provide systems-theory precedents (Redheffer, 1960; Hanzon et al., 2006). The present contribution connects these structures to bounded associative content, its explicit nonlinear remainder, shared-coordinate expressivity and exact chunk evaluation.

# B Geometric details

## B.1 Write algebra

For $\kappa > 0$, set $\widehat{F}(M) = \sqrt{\kappa} F(M/\sqrt{\kappa})$ and use the orthonormal bases and blocks of Section 3. Then $P = \operatorname{diag}(c, I)$ and $Q = \operatorname{diag}(c, I)$ in their respective spaces. Inverting the block-upper-triangular denominator gives the stated blocks with $\tau = \sqrt{\kappa}\|u\|\|k\|/c$ and $h = c^{-1}$; the tangent's left and right factors are $\operatorname{diag}(h, I)$. A zero value or key gives identity. The sequential pivot is

$$\delta = 1 + \frac{\kappa}{c} u^\top X k > 0, \qquad \left|\frac{\kappa}{c} u^\top X k\right| \le \frac{\sqrt{\kappa}\|u\|\|k\|}{c} < 1, \tag{9}$$

with $\delta = 1$ at zero controls. All identities and invariance claims use exact arithmetic.

A raw write $u = \epsilon v$ has Riccati expansion

$$F_{\kappa,\epsilon v,k}(X) = X + \epsilon(vk^\top - \kappa X k v^\top X) + O(\epsilon^2). \tag{10}$$

The time-$t$ flow of this vector field is (2) with $u = \sinh(\sqrt{\kappa}\|v\|\|k\| t)v/(\sqrt{\kappa}\|v\|\|k\|)$, taking continuous zero limits. Its finite-write rapidity is $\operatorname{asinh}(\sqrt{\kappa}\|u\|\|k\|)$. Decay and erasure are separate contractions.

## B.2 Translation coordinates

Fix oriented unit directions and their complements. For $u = \lambda\widehat{u}$, $k = \mu\widehat{k}$ and $\kappa > 0$, put $t = \sqrt{\kappa}\lambda\mu$, $\tau = t/\sqrt{1+t^2}$, $h = (1+t^2)^{-1/2}$ and $\theta = \operatorname{asinh} t = \operatorname{atanh}\tau$. The block formula holds for signed $\tau$, including zero. Define

$$\mathcal{C}(M) = (\xi, \pi, \varrho, C), \qquad \xi = \operatorname{atanh} x, \quad \pi = \frac{p_\perp}{\sqrt{1-x^2}}, \quad \varrho = \frac{r_\perp}{\sqrt{1-x^2}}, \quad C = Z + x\pi\varrho^\top.$$

Its $VK$ coordinates have inverse $x = \tanh\xi$, $p_\perp = \operatorname{sech}\xi\,\pi$, $r_\perp = \operatorname{sech}\xi\,\varrho$, $Z = C - \tanh\xi\,\pi\varrho^\top$. Since $|x| < 1$ and $1 + \tau x > 0$,

$$x' = \tanh(\xi + \theta), \qquad 1 - x'^2 = \frac{h^2(1-x^2)}{(1+\tau x)^2}, \qquad x' - x = \frac{\tau(1-x^2)}{1+\tau x}.$$

Thus $\pi' = \pi$, $\varrho' = \varrho$ and $Z' = Z - (x' - x)\pi\varrho^\top$, giving $C' = C$. Applying rapidity $-\xi$ and its inverse shows that $\mathcal{C}$ is a smooth bijection onto

$$\mathbb{R} \times \mathcal{D}, \qquad \mathcal{D} = \left\{(\pi, \varrho, C) : \left\|\begin{pmatrix} 0 & \varrho^\top \\ \pi & C \end{pmatrix}\right\| < 1\right\}.$$

Every pure write therefore translates only $\xi$ by $\theta$ in these common coordinates. Empty blocks cover dimension one; at $\kappa = 0$, the writes are already translations.

### B.3 UNIFORM RADIUS

Suppose $\|A_t\| \le \rho < 1$ and $\|u_t\|\|k_t\| \le B$ uniformly.

**Proposition 2** (Uniform radius). *For fixed $\kappa \ge 0$ and $S_0 = 0$, every state in (1) satisfies*

$$\|S_t\| \le R_\kappa := \frac{2B}{c_B(1-\rho) + \sqrt{c_B^2(1-\rho)^2 + 4\rho\kappa B^2}}, \qquad c_B = \sqrt{1+\kappa B^2}. \tag{11}$$

*For $B > 0$, $R_0 = B/(1-\rho)$ and $0 < R_\kappa < 1/\sqrt{\kappa}$ for $\kappa > 0$. The radius is sharp over this abstract control class.*

*Proof.* For $\kappa > 0$, set $s = \sqrt{\kappa}\|S\|$ and $w_B = \sqrt{\kappa}B/c_B$. The classical ball distance satisfies $d(0, M) = \operatorname{atanh}\|M\|$. Its triangle inequality and invariance under the write give $s_t \le (\rho s_{t-1} + w_B)/(1 + \rho w_B s_{t-1})$. The increasing scalar map has fixed point $\sqrt{\kappa}R_\kappa < 1$; solving its quadratic gives (11). Aligned rank-one controls with $A_t = \rho I$ attain this comparison and approach the fixed point. For $\kappa = 0$, iterate $\|S_t\| \le \rho\|S_{t-1}\| + B$. □

For the affine tangent, fixed $\kappa > 0$ and $\|A_t\| \le \rho < 1$ give the sharp dimension-free spectral bound from $S_0 = 0$, for all finite writes,

$$\|S_t\| \le \frac{\mathcal{R}_\rho}{\sqrt{\kappa}}, \qquad \mathcal{R}_\rho = \begin{cases} 1, & 0 \le \rho \le \frac{1}{2}, \\ [4\rho(1-\rho)]^{-1/2}, & \frac{1}{2} < \rho < 1. \end{cases} \tag{12}$$

Over abstract controls with $\|A\| \le \rho \le 1$, the tangent preserves the prescribed open ball $\mathcal{B}_\kappa$ for every finite write if and only if $\rho \le \frac{1}{2}$; the curved write does so for every $\rho \le 1$. The proof below uses independently chosen controls.

**Frozen frontend bounds.** Pre-RMSNorm gives $\|z\| \le R_{\text{in}} = \sqrt{d}\|\gamma\|_\infty$. For row $f_i$ of the two decay projections, head log-rate $\alpha_h$ and offset $\mathrm{dt}_i$,

$$d_{t,i} \le \rho_i := \exp\{-e^{\alpha_h}\,\mathrm{softplus}(\mathrm{dt}_i - R_{\text{in}}\|f_i\|)\} < 1.$$

Thus $\rho = \max_i \rho_i < 1$. For convolution weights $c_{j,s}$, bias $b_j$, value-projection row $W_{v,j}$ and zero initial history, $|\mathrm{SiLU}(x)| \le |x|$ gives

$$|v_{t,j}| \le L_j := R_{\text{in}}\|W_{v,j}\| \sum_s |c_{j,s}| + |b_j|, \qquad \|u_{t,h}\| \le \Big(\sum_{j\in h} L_j^2\Big)^{1/2}.$$

Sigmoid writes and subunit keys supply finite $B$ for either signed or positive erasure. For admissible $S_0$, $\|S_t\| \le \max\{\|S_0\|, R_\kappa\}$; when $B = 0$, $\|S_t\| \le \rho^t\|S_0\|$.

### B.4 SHARP TANGENT SPECTRAL RADIUS

For a nonzero write put $p = u/\|u\|$, $q = k/\|k\|$, $\eta = 1/c$ and $\omega = \sqrt{1-\eta^2} < 1$. In normalized coordinates $M = \sqrt{\kappa}S$,

$$T(M) = LMR + \omega pq^\top, \quad L = I - (1-\eta)pp^\top, \quad R = I - (1-\eta)qq^\top.$$

Zero values or keys give identity, covered by $\omega = 0$. For $\|M\| \le r$ and unit $x, y$, put $h = |x^\top p|$, $j = |q^\top y|$. Since $L^2 = I - \omega^2 pp^\top$ and $R^2 = I - \omega^2 qq^\top$, arithmetic–geometric mean on both products gives

$$\begin{aligned} |x^\top T(M)y| &\le r\sqrt{1-\omega^2h^2}\sqrt{1-\omega^2j^2} + \omega hj \\ &\le r + (\omega - r\omega^2)(h^2 + j^2)/2 \le \max\{r, r(1-\omega^2) + \omega\}. \end{aligned}$$

Taking the supremum over $0 \le \omega < 1$ gives $\|T(M)\| \le \phi(r)$, where $\phi(r) = 1$ for $r \le \frac{1}{2}$ and $\phi(r) = r + 1/(4r)$ otherwise. This nondecreasing map satisfies $\phi(\rho\mathcal{R}_\rho) = \mathcal{R}_\rho$, proving (12). Its 1-Lipschitz property also gives $\|S_t\| \le \mathcal{R}_\rho/\sqrt{\kappa} + \rho^t(\|S_0\| - \mathcal{R}_\rho/\sqrt{\kappa})_+$ for finite $S_0$ ($\rho^0 = 1$). For $\rho \le \frac{1}{2}$, $\omega \uparrow 1$ approaches the bound in one write. For $\frac{1}{2} < \rho < 1$, fixed aligned $p, q$, $A_t = \rho I$ and finite $\omega_* = \sqrt{(1-\rho)/\rho} < 1$ give $M_t = \mathcal{R}_\rho[1 - (2\rho-1)^t]pq^\top \to \mathcal{R}_\rho pq^\top$. If $\rho \le \frac{1}{2}$, the displayed maximum is strictly below one for $r \le \rho$ and every finite write. If $\rho > \frac{1}{2}$, an admissible strict contraction above one half gives an aligned trajectory that leaves even the closed unit ball, proving the exact open-ball threshold.

### B.5 Retention and propagation

On the unit matrix ball, use the classical invariant tangent norm and its path distance $d$ (Harris, 1974; Ostrovskii et al., 2009),

$$\mathfrak{k}_M(H) = \|(I - MM^\top)^{-1/2} H (I - M^\top M)^{-1/2}\|. \tag{13}$$

**Proposition 3** (Fixed-control retention). *For $\kappa > 0$, two curved trajectories $S_t, S'_t$ with identical controls and normalized states $M_t = \sqrt{\kappa} S_t$, $N_t = \sqrt{\kappa} S'_t$ obey*

$$d(M_t, N_t) \le \prod_{j=1}^{t} \|A_j\|\, d(M_0, N_0), \tag{14}$$

*with sharp one-step coefficient $r_t = \|A_t\|$. Pure writes preserve $d$ exactly. At $\kappa = 0$, the same factor bounds ordinary spectral-norm differences.*

A write is the ball automorphism

$$\Phi_W(P) = (I - WW^\top)^{-1/2}(P + W)(I + W^\top P)^{-1}(I - W^\top W)^{1/2},$$

which is an isometry of (13). Classical Schwarz–Pick makes $M \mapsto MQ$ nonexpansive when $\|Q\| \le 1$. For $0 \le r \le 1$,

$$\mathfrak{k}_{rM}(rH) = r\|L_M[(I - MM^\top)^{-1/2} H (I - M^\top M)^{-1/2}] R_M\| \le r\mathfrak{k}_M(H),$$

where $L_M = (I - r^2 MM^\top)^{-1/2}(I - MM^\top)^{1/2}$ and the analogous right factor $R_M$ are contractions by functional calculus. Factor right multiplication by $A$ as $M \mapsto M(A/r) \mapsto rM$ when $r = \|A\| > 0$; $r = 0$ is constant. Integrating proves (14). At zero, choose a unit rank-one $H$ with $\|HA\| = r\|H\|$. The write isometry preserves this tangent ratio, proving sharpness. For the affine boundary, differences obey $\Delta_t = \Delta_{t-1} A_t$ directly.

**Propagation of the exact remainder.** Freeze the controls on a curved trajectory and initialize a tangent trajectory at the same state. Write $T_t(X) = L_t X R_t + B_t$, $B_t = w_t u_t k_t^\top$, $E_t = S_t - \widetilde{S}_t$, and $\Delta_t = F_t(S_{t-1}A_t) - T_t(S_{t-1}A_t)$. Then

$$\widetilde{S}_t = L_t \widetilde{S}_{t-1} A_t R_t + B_t, \qquad E_t = L_t E_{t-1} A_t R_t + \Delta_t. \tag{15}$$

Since $\|L_t\|, \|R_t\| \le 1$ and $E_0 = 0$, iteration gives

$$\|E_t\| \le \sum_{i=1}^{t} \|\Delta_i\| \prod_{j=i+1}^{t} \|A_j\|. \tag{16}$$

### B.6 Closed-write algebra and affine obstructions

Fix a unit key, $D = I, e = 0$, $\kappa > 0$, and $V \ge 2$. For values $u, v$, put

$$c_u = \sqrt{1 + \kappa\|u\|^2}, \quad P_u = I + \frac{\kappa u u^\top}{c_u + 1}, \quad \begin{pmatrix} A_* & B_* \\ C_* & c_* \end{pmatrix} = \begin{pmatrix} P_v & v \\ \kappa v^\top & c_v \end{pmatrix} \begin{pmatrix} P_u & u \\ \kappa u^\top & c_u \end{pmatrix}. \tag{17}$$

**Proposition 4** (Closed-write transport). *Writes $u$, $v$, then $-B_*$ at that key compose exactly to*

$$S \longmapsto \mathcal{R}S, \qquad \mathcal{R} = A_* - \frac{B_* C_*}{c_* + 1} \in SO(V). \tag{18}$$

*For linearly independent $u, v$, this is a nonidentity, nonscalar rotation. It fixes zero and preserves every singular value of every state. No fixed-control sequence $S \mapsto SA_t + B_t$ realizes this same operation on the whole open ball.*

Write $T(u) = \left(\begin{smallmatrix} P_u & u \\ \kappa u^\top & c_u \end{smallmatrix}\right)$. It preserves $J_\kappa = \mathrm{diag}(\kappa I_V, -1)$, has determinant one, and is identity on unused key directions. For $T(v)T(u)$ in (17),

$$A_* = P_v P_u + \kappa v u^\top, \quad B_* = P_v u + c_u v, \quad C_* = \kappa(v^\top P_u + c_v u^\top), \quad c_* = c_v c_u + \kappa v^\top u > 0.$$

The preserved form implies $c_*^2 - \kappa\|B_*\|^2 = 1$ and $\kappa B_*^\top A_* = c_* C_*$. Hence $c_{B_*} = c_*$ and direct multiplication gives $T(-B_*)T(v)T(u) = \mathrm{diag}(\mathcal{R}, 1)$ with (18). Form preservation and determinant one imply $\mathcal{R} \in SO(V)$. For independent $u, v$,

$$A_* - A_*^\top = \kappa \left(1 + \frac{\kappa u^\top v}{(c_u + 1)(c_v + 1)}\right)(vu^\top - uv^\top) \neq 0.$$

If $\mathcal{R} = I$, $T(v)T(u) = T(B_*)$ would have symmetric upper block, a contradiction. Nor can $\mathcal{R} = -I$. Conjugating each boost by $\mathrm{diag}(\sqrt{\kappa}I, 1)$ makes it positive definite, so their product has positive eigenvalues; $T(B_*)\,\mathrm{diag}(-I, 1)$ instead has eigenvalue $-1$ on a nonzero value direction perpendicular to $B_*$. Collinear values give identity.

At fixed $u, v$, $B_* \to u + v$ and $\mathcal{R} = I + \kappa\Omega + O(\kappa^2)$, where $\Omega = (vu^\top - uv^\top)/2$. Two same-key writes closing zero are inverses, so three writes are minimal.

**Strict-control separation.** Let $d = \min_\lambda \|\mathcal{R} - \lambda I\| > 0$ and $\mathcal{K}_r = \{S : \|S\|_F \leq r\}$, $0 < r < 1/\sqrt{\kappa}$. Centered rank-one tests $S = xz^\top$ eliminate offsets and show $\inf_{A,B} \sup_{\mathcal{K}_r} \|\mathcal{R}S - SA - B\|_F = rd$; the scalar minimizer attains equality. Choose a small nonzero loop with angle $|\theta| < \pi/2$, so $d = |\sin\theta|$. Preceding each write by $A_\tau = (1 - \tau)(I - \tau kk^\top)$ gives strict controls and a composition $G_\tau$ with $\varepsilon_\tau := \sup_{\mathcal{K}_r} \|G_\tau(S) - \mathcal{R}S\|_F \to 0$. Consequently

$$\inf_{A,B} \sup_{S \in \mathcal{K}_r} \|G_\tau(S) - (SA + B)\|_F \geq rd - \varepsilon_\tau > 0 \tag{19}$$

for small $\tau > 0$.

A finite tangent sequence is $\Phi(S) = PSQ + B$, where $P = \prod_{j=m}^{1} L_j$, $Q = \prod_{j=1}^{m}(A_j R_j)$ and $L_j = I - (1 - \eta_j)p_j p_j^\top$, $\eta_j = 1/c_j \in (0, 1]$. All factors contract, including signed or singular symmetric erasers; zero write vectors give identity factors. Hence $\|P\|, \|Q\| \leq 1$ and $\|P - I\| \leq \sum_j (1 - \eta_j) \leq -\log\det P$. For a unit query $q$, set $E_q = \sup_{\mathcal{K}_r} \|(\mathcal{R}S - \Phi(S))q\|$ and $\delta = E_q/r$. Centered tests $S = xq^\top$ give $\|\mathcal{R} - \alpha P\| \leq \delta$, $\alpha = q^\top Q q$, $|\alpha| \leq 1$. If $\delta < 1$, then $\alpha \neq 0$, $\sigma_{\min}(P) \geq 1 - \delta$ and $\det P \geq (1 - \delta)^V$. Thus

$$d \leq \delta + |\alpha|\|P - I\| \leq \delta - V\log(1 - \delta) \leq \frac{(V + 1)\delta}{1 - \delta}, \qquad E_q \geq \frac{rd}{V + 1 + d}.$$

The last bound also holds for $\delta \geq 1$. Replacing the loop by $G_\tau$ reduces it by at most $\varepsilon_\tau$, uniformly over queries, sequence lengths, offsets and finite nonnegative comparator curvatures.

The sharp small-curvature separation is

$$\inf_{\Phi} \sup_{\|S\|_F \leq r} \|(\mathcal{R}_\kappa S - \Phi(S))q\| = r\kappa\|\Omega\| + O(\kappa^2), \qquad \kappa \downarrow 0. \tag{20}$$

Indeed, put $h_j = 1 - \eta_j$ and $\ell = \sum_j h_j \leq t = -\log\det P$. Product expansion gives $P = I - \sum_j h_j p_j p_j^\top + E$ with $\|E\| \leq e^\ell - 1 - \ell$. Its first-order term is symmetric, so

$$\|\mathrm{skew}\, P\| \leq e^t - 1 - t, \qquad \|\mathrm{skew}\,\mathcal{R}\| \leq \delta + (1 - \delta)^{-V} - 1 + V\log(1 - \delta) = h_V(\delta).$$

For $\delta < 1$, this uses $t \leq -V\log(1 - \delta)$ and $\|\mathcal{R} - \alpha P\| \leq \delta$. Since $h_V$ is increasing and $h_V(\delta) = \delta + O(\delta^2)$, the lower bound and identity comparator prove (20). Strict perturbations of size $o(\kappa)$ preserve its leading coefficient.

These separations allow independent controls; a fixed retention bound need not permit $\tau \downarrow 0$.

**Proof of the finite affine gap.** Normalize $X = \sqrt{\kappa}S$ at fixed $\kappa > 0$, and put $\mathbb{B} = \{X \in \mathbb{R}^{V \times K} : \|X\| < 1\}$ and $\mathbb{B}_r = \{X : \|X\| \leq r\}$; these are spectral balls, distinct from $\mathcal{K}_r$ above. With coordinate vectors $e_i, f_a$, define

$$B_{ia} = \begin{pmatrix} 0 & e_i f_a^\top \\ f_a e_i^\top & 0 \end{pmatrix}, \qquad g_{ia}(t) = \exp(tB_{ia}).$$

For $g = \left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$, $g \cdot X = (AX + B)(CX + D)^{-1}$. Each $g_{ia}(t)$ is exactly the normalized pure write (2) with $k = f_a$, $u = \sinh(t)e_i/\sqrt{\kappa}$; its inverse negates $u$. The same map has any prescribed subunit key norm $\sigma \in (0, 1)$ by taking $k = \sigma f_a$ and $u = \sinh(t)e_i/(\sigma\sqrt{\kappa})$.

We prove Proposition 1 with these definitions; the coordinate writes of the proposition are the maps $g_{ia}(t)$.

*Proof of Proposition 1.* We first exclude a local affine realization of all small coordinate boosts. Suppose $F(g_j(t)\cdot X) = A_j(t)F(X)$ near $X = 0, t = 0$, enumerating the $B_{ia}$ by $j$. Invariance of domain (Hatcher, 2002, Theorem 2B.3) makes the image open. Values at $n+1$ affinely independent image points uniquely determine $A_j(t)$ and give continuous coefficients. On a smaller common overlap, the boost identities force $A_j(s+t) = A_j(s)A_j(t)$ and inverses, first on an open image and hence globally as affine maps. Extending by subdivision gives continuous one-parameter groups, which are smooth by Conrad & Landesman (2018, Theorem 6.16) and hence affine matrix exponentials. The analytic chart $\chi(t) = g_1(t_1)\cdots g_n(t_n)\cdot 0$ has derivative basis $e_i f_a^\top$ at zero. Keeping all intermediate points in the common neighborhood, $F(\chi(t)) = A_1(t_1)\cdots A_n(t_n)F(0)$ shows that $F$ is analytic there. Differentiating the same word identity at $X = 0$ gives $DF_{\chi(t)}D(g_1(t_1)\cdots g_n(t_n))_0 = L_t DF_0$ with invertible word derivatives and affine linear parts $L_t$. Thus $DF$ has constant rank near zero; injectivity and the constant-rank theorem force rank $n$. Consequently $F$ is a local diffeomorphism, without any assumed smoothness.

For $M = \left(\begin{smallmatrix} P & Q \\ Q^\top & R \end{smallmatrix}\right) \in \mathfrak{so}(V,K)$, with $P, R$ skew, the fundamental field is $\xi_M(X) = PX + Q - XR - XQ^\top X$. Coordinate boosts generate this Lie algebra,

$$[B_{ia}, B_{jb}] = \operatorname{diag}\big(\delta_{ab}(e_i e_j^\top - e_j e_i^\top), \delta_{ij}(f_a f_b^\top - f_b f_a^\top)\big).$$

Their pushed-forward fields are affine. Brackets and linear combinations remain affine; every relation holds globally because an affine field vanishing on an open set is zero. With the same fundamental-field sign convention on both sides, this gives a Lie-algebra action into $\mathfrak{aff}(n)$. For $V \geq 2, K \geq 1$, $\mathfrak{so}(V,K)$ is semisimple, including $\mathfrak{so}(2,1)$ and $\mathfrak{so}(2,2)$. Homogenize the affine action on $\mathbb{R}^{n+1}$. Weyl complete reducibility (Etingof, 2022, Section 18.4) gives an invariant line complementary to $\mathbb{R}^n \times \{0\}$. The quotient action is trivial, so its vector $(o, 1)$ is annihilated by every generator: $o$ is a common affine fixed point. Translate $o$ to zero and write $p = F(0)$. All linear isotropy fields fix $p$. If $p \neq 0$, nearby distinct points $\lambda p$ also lie in the open local image and are fixed by these fields. Their preimages satisfy $PX - XR = 0$ for every skew $P, R$; already $PX = 0$ for all skew $P$ forces $X = 0$ when $V \geq 2$, contradicting injectivity. Hence $p = 0$. But every linear boost field then vanishes at $p$, contradicting $DF_0(e_i f_a^\top) \neq 0$.

It remains to force this local realization from finitely many controls. Set $r_0 = r/2$ and choose $h > 0$ such that $g_j(t)\mathbb{B}_{r_0} \subset \operatorname{int}\mathbb{B}_r$ for all $j$ and $|t| \leq h$, using compactness. Set $\tau_1 = h/3$, $\tau_2 = h/(3\sqrt{2})$ and $\mathcal{S}_r = \{g_j(\pm\tau_1), g_j(\pm\tau_2) : 1 \leq j \leq n\}$. If $\delta_{F,r} = 0$, affine approximants converge at an affine basis in $F(\operatorname{int}\mathbb{B}_{r_0})$; interpolation forces convergence of all coefficients to exact maps $\Psi_s$ on $F(\mathbb{B}_r)$. Inverse pairs compose to identity on the smaller open image, hence globally.

For $q = m\tau_1 + \ell\tau_2 \in [-h, h]$, reorder its signed steps so every partial sum stays in $[-h, h]$. From a positive sum choose a negative step, and otherwise a positive step, while available. Opposite-sign steps have size at most $h$; if that sign is exhausted, the remainder moves monotonically to $q$. Thus every intermediate state from $\mathbb{B}_{r_0}$ stays in $\mathbb{B}_r$, and the exact affine word agrees with $F(g_j(q)\cdot X)$. Irrationality of $\tau_1/\tau_2$ supplies such $q$ tending to each $|t| < h/2$. Interpolation at the same affine basis and continuity give exact affine maps for every small $g_j(t)$ on the common ball $\mathbb{B}_{r_0}$, contradicting the local result. Here the zero-error limit precedes the finite words. □

Preceding each $s \in \mathcal{S}_r$ by $A_\tau = (1-\tau)(I - \tau k k^\top)$ gives targets converging uniformly to $s \cdot X$ on $\mathbb{B}_r$. Their images remain in a compact subset of $\mathbb{B}$, so uniform continuity preserves a gap of at least $\delta_{F,r}/2$ for sufficiently small $\tau > 0$. The gap depends on $F$ and scales with it. In the excluded scalar case, $\operatorname{atanh} X$ conjugates boosts to translations. A larger frame $(P, Q)$ updates linearly and reads $PQ^{-1}$, but uses $(V+K)K$ coordinates.

**Normalized reads.** The following states have $p_\perp = 0$. Their unchanged $Z$ isolates the state-dependent gain on $r_\perp$, before learned gains and projections.

**Proposition 5** (Eight-state quantitative normalized gap)**.** *Fix $V, K \geq 2$, $0 < r < 1$, $t > 0$, $c = \cosh t$, $s = \sinh t$, $\zeta = r/4$, $b = s\zeta$, and $F = F_{1,se_1,f_1}$. Use the eight states $X = \zeta(Pe_1 f_1^\top + Ae_1 f_2^\top + Ze_2 f_2^\top)$ with $A \in \{1,2\}$, $P, Z \in \{-1, 1\}$. For $N_\epsilon(y) = y/\sqrt{\|y\|^2/V + \epsilon}$, $\epsilon \geq 0$, and $N_\epsilon(0) = 0$, define*

$$E_\epsilon = \inf_{\Psi \text{ affine}} \max_{A,P,Z} \|N_\epsilon(\Psi(X)f_2) - N_\epsilon(F(X)f_2)\|.$$

*Over all $\Psi : \mathbb{R}^{V\times K} \to \mathbb{R}^{V\times K}$, with no coefficient, magnitude or sequence-length bound,*

$$\frac{\sqrt{V}b}{32c^2} \le \delta_8 := E_0 \le \frac{3\sqrt{V}b}{c^2}, \qquad E_\epsilon \ge \frac{m_\epsilon \delta_8}{2\sqrt{V}} > 0,$$

*where $\gamma = \zeta\sqrt{1+(c+b)^{-2}}$ and $m_\epsilon = \sqrt{V}\gamma/\sqrt{\gamma^2+V\epsilon}$.*

*Proof.* All states have $\|X\| \le \sqrt{6}\zeta < r$, and $0 < b/c < 1/4$. Substitution gives $y = F(X)f_2 = \zeta[A/(c+bP), Z, 0, \dots]^\top$, with direction $Q = (A, (c+bP)Z, 0, \dots)^\top$. For a comparator of maximum error $\sqrt{V}\tau$, the lower bound is immediate if $\tau \ge 1$. Otherwise every query is nonzero; positively scale its affine query $\ell$ so $\max\|\ell\| = 1$, preserving $N_0$. Write $\ell_1 = U + WA + BP + CZ$, $\ell_2 = D + JA + EP + TZ$. For $g := A\ell_2 - (c+bP)Z\ell_1$, the error gives $|g| \le R := \tau\sqrt{4+(c+b)^2}$ and $\|\ell_\perp\| \le \tau$ for all remaining coordinates. The odd/even parts in $Z$ of $g$ are $ZG, H$, where

$$G = A(T - cW) - (cU + bB) - P(cB + bU + bWA), \qquad H = A(D + JA) - cC + P(AE - bC).$$

Both have magnitude at most $R$. Differences and averages at $A = 1, 2$, $P = \pm 1$ give

$$|W| \le 2R/b, \quad |T - cW| \le 2R, \quad |cB + bU|, |cU + bB| \le 3R,$$
$$|E| \le 2R, \quad |C| \le 3R/b, \quad |D + JA| \le 3cR/b + R.$$

Thus $|U|, |B| \le 4R/c$, $|T| \le 2cR/b + 2R$, and at every witness $|\ell_1| \le 9R/b$, $|\ell_2| \le 25cR/(4b)$, so $\|(\ell_1, \ell_2)\| \le 11cR/b$. Since $R \le 5c\tau/2$, the norm-one witness gives $1 \le \tau^2[1 + (55c^2/(2b))^2] \le (32c^2\tau/b)^2$, using $b^2 \le c^4$ and $1 + (55/2)^2 < 32^2$. This proves the lower bound, including zero queries and extra dimensions.

For the upper bound use the affine query $\ell = (A, cZ, 0, \dots)^\top$. The horizontal coordinates $A/(c+bP)$ and $A/c$ differ by at most $2b/[c(c-b)] \le 8b/(3c^2)$. The derivative norm of $x \mapsto (x, Z)/\sqrt{x^2+1}$ is at most one, proving the stronger upper bound $8\sqrt{V}b/(3c^2)$. Finally $\min\|y\| = \gamma$ at $A = P = 1$, so $\min\|N_\epsilon(y)\| = m_\epsilon$. For nonzero $a, b$, $\|a/\|a\| - b/\|b\|\| \le 2\|a-b\|/\|b\|$ bounds zero-epsilon error by $2\sqrt{V}/m_\epsilon$ times epsilon error; zero queries also satisfy this. Take maxima and infima. □

Fixed-control affine and tangent sequences are included because they compose to affine state maps. The result concerns canonical normalized reads.

### B.7 Native oblique erasure

Let $F_\kappa$ be a holomorphic self-map of the complex spectral ball of radius $\kappa^{-1/2}$ and $Z = \sqrt{\kappa}F_\kappa(0)$. Schwarz–Pick at zero gives

$$\|(I - ZZ^*)^{-1/2} DF_\kappa(0)[H](I - Z^*Z)^{-1/2}\| \le \|H\|.$$

The positive square-root contractions imply $\|DF_\kappa(0)\| \le 1$. Derivative convergence to $F_0(S) = SA + U$ therefore requires $\|A\| \le 1$, with $\|DF_\kappa(0) - R_A\| \ge \max(0, \|A\| - 1)$ for $R_A(H) = HA$. Native GDN2 permits $k = (1,1)^\top/\sqrt{2}$, $b = (1/2, 3/2)$ and $D = (19/20)I$, giving $E = I - \mathrm{diag}(b)kk^\top = \frac{1}{4}\left(\begin{smallmatrix} 3 & -1 \\ -3 & 1 \end{smallmatrix}\right)$ and $\|DE\| = 19\sqrt{5}/40 > 1$, although $E^2 = E$.

## C Complete erase-eliminated chunk

The fractional write has the rank-one scattering form

$$F(X) = X + (wu - \alpha Xk)k^\top - (\beta u + gXk)\frac{u^\top X - \alpha(u^\top Xk)k^\top}{1 + gu^\top Xk}, \tag{21}$$

with coefficients (3).

**Proposition 6** (Exact chunk with one unknown per token)**.** *For an admissible initial state and split controls, (7) with (23) returns every post-write read and the exact final sequential state. The leading minors of $H$ are $\det H_{1:r,1:r} = \prod_{i=1}^{r}\delta_i > 0$, so unpivoted LU exists. At fixed $\kappa = 0$, $H = K_s = I$ and no dense LU is needed.*

Rows of $U, E, K, Q$ are the token vectors $u_i^\top, e_i^\top, k_i^\top, q_i^\top$. Use componentwise decay products

$$I_{ij} = \prod_{i<\ell\le j} d_\ell \quad (i \le j), \qquad p_i = \prod_{\ell\le i} d_\ell, \qquad t_i = I_{i,n}.$$

They are evaluated directly, never as reciprocal prefix ratios. Let $\ell_{ij} = 1\{i < j\}$ and $\tau_{ij} = 1\{i \le j\}$. Construct

$$(J_{EE})_{ij} = \ell_{ij} \sum_a e_{ia} I_{ij,a} e_{ja}, \qquad (J_{WE})_{ij} = \ell_{ij} \sum_a k_{ia} I_{ij,a} e_{ja},$$
$$(J_{EW})_{ij} = \tau_{ij} \sum_a e_{ia} I_{ij,a} k_{ja}, \qquad (J_{WW})_{ij} = \ell_{ij} \sum_a k_{ia} I_{ij,a} k_{ja},$$
$$(Z_E)_{ij} = \tau_{ij} \sum_a e_{ia} I_{ij,a} q_{ja}, \qquad (Z_W)_{ij} = \tau_{ij} \sum_a k_{ia} I_{ij,a} q_{ja}.$$

The diagonal of $J_{EW}$ is included because erasure precedes writing. Set

$$M_E = S_0(p \odot E)^\top, \qquad M_W = S_0(p \odot K)^\top,$$
$$R_E = [Z_E, E \odot t], \qquad R_W = [Z_W, K \odot t],$$
$$T_0 = [S_0(p \odot Q)^\top, S_0 \operatorname{diag}(p_n)].$$

A single unit upper-triangular solve with concatenated right-hand sides gives

$$\begin{aligned} (I + J_{EE})[P_E, Z] &= [J_{EW}, R_E], \\ \bar{M} = M_W - M_E P_E, \quad \bar{J} &= J_{WW} - J_{WE} P_E, \\ \bar{R} = R_W - J_{WE} Z, \quad \bar{T} &= T_0 - M_E Z. \end{aligned} \tag{22}$$

Here $P_E$ is upper triangular and $\bar{J}$ strictly upper triangular. Let $G = \mathrm{sl}(UU^\top)$ be the strict lower part of the value Gram matrix, $\Omega = U\bar{M}$, and $[z] = \operatorname{diag}(z)$. The reduced system and its reconstruction are

$$\begin{aligned} A_v &= I + G[\beta] + \Omega[g], & B_v &= -G[w] + \Omega[\alpha], \\ K_s &= I + \bar{J}[\alpha], & K_s[Z_d, z_0] &= [\bar{J}[g], \bar{R}], \\ H &= A_v - B_v Z_d, & R_* &= U\bar{T} - B_v z_0, \\ z_k &= z_0 - Z_d z_v, & [Y^\top, S_n] &= \bar{T} + U^\top(-[\beta] z_v + [w] z_k) - \bar{M}([g] z_v + [\alpha] z_k), \end{aligned} \tag{23}$$

with $z_v$ from (7). The first $n$ columns of $[Y^\top, S_n]$ are the reads and the last $K$ columns the final state. The dense system has size $n$ instead of $2n$, which reduces its leading cubic LU work eightfold and interval storage fourfold.

**Derivation and causality.** For positive decays, substitute $S_i = N_i \operatorname{diag}(p_i)$ into the sequential scattering recurrence. Its left/right key factors are $\operatorname{diag}(p_i)k_i$ and $\operatorname{diag}(p_i)^{-1}k_i$. Each interacting pair at indices $i < j$ cancels the prefixes into $I_{ij}$; query pairs use $i \le j$. This yields the generic event value/key linear system. Erase value rows have zero right-hand side and are identity rows. Their key rows are

$$(I + J_{EE}) z_{k,E} + J_{EW}([g] z_v + [\alpha] z_k) = R_E.$$

Substituting $z_{k,E} = Z - P_E([g] z_v + [\alpha] z_k)$ produces (22). The remaining value/key equations are $A_v z_v + B_v z_k = U\bar{T}$ and $\bar{J}[g] z_v + K_s z_k = \bar{R}$, whose Schur elimination gives (7) and (23). In the full two-event value Schur matrix, erase rows and columns are isolated identity entries, since erase values vanish and their $\beta, g$ are zero. Its unpivoted factors therefore restrict to the write-index factors in (7). Restricting the original causal mask to after-write queries gives exactly $\tau$. Upper-triangular erase elimination preserves each prefix; middle masking after the lower solve computes that prefix's solve, and the upper solve preserves its zero tail. Hence all query columns, as well as the final state, are exact.

Sequential elimination gives write pivots (9) and erase pivots one, so leading determinants are their products. Entries are continuous polynomials in decay coordinates. Taking zero-decay limits preserves the identities where sequential pivots remain nonzero; no reciprocal decay is evaluated. At fixed $\kappa = 0$, $\beta = \alpha = g = 0, w = 1$, giving $H = K_s = I$ and $[Y^\top, S_n] = \bar{T} + U^\top \bar{R}$.

### C.1 UNSCALED INVERSE AND CONDITION BOUNDS

**Proposition 7** (Interior conditioning of the implemented system). *Fix $\kappa > 0$, $\|u_i\| \le U_* < \infty$ and $\|k_i\| \le K_* \le 1$. Suppose every normalized state before and after writes and contractions satisfies $\|M\| \le s < 1$. For $n$ tokens, including zero-vector ports, the unscaled matrix $H$ satisfies*

$$\|H^{-1}\| \le C_n, \qquad \|H\| \le B_n, \qquad \mathrm{cond}_2(H) \le B_n C_n = O(n^4), \tag{24}$$

*where the constants in (25) depend only on $s$ and $t_* = \sqrt{\kappa} U_* K_*$, with $B_n, C_n = O(n^2)$. At $\kappa = 0$, $H = I$ exactly.*

$$\begin{aligned} \nu &= (1-s^2)^{-1/2}, \qquad a_* = (1-s)^{-1}, \qquad t_* = \sqrt{\kappa} U_* K_*, \\ C_n &= 1 + n\nu t_* \big[ (t_*/2 + s a_*)(1 + n\nu s a_*) + s a_* \big], \\ B_n &= 1 + n(t_*^2/2 + s t_*) + n^2 t_*^2 (1 + s t_*/2). \end{aligned} \tag{25}$$

**The reduced system as a transfer network.** Combine the right contractions before write $i$ into $Q_i$, preserving their order; for the native split $Q_i = D_i(I - e_i e_i^\top)$. The erase-eliminated quantities in (22) have the ordered-product expressions

$$\bar{M}_{:j} = S_0 Q_1 \cdots Q_j k_j, \qquad \bar{J}_{ij} = 1\{i < j\} k_i^\top Q_{i+1} \cdots Q_j k_j.$$

With $C = \bar{J}[g]$, the full system is $\mathcal{F} = \left(\begin{smallmatrix} A_v & B_v \\ C & K_s \end{smallmatrix}\right)$ and $H = A_v - B_v K_s^{-1} C$. Since $K_s$ is unit upper triangular, $H^{-1}$ is obtained by forcing the value ports, $\mathcal{F}(z_v, z_k)^\top = (b, 0)^\top$.

Use normalized transfer coordinates $\widehat{x} = \sqrt{\kappa} x_{\mathrm{phys}}$. At each write, minus and plus denote its adjacent input and output boundaries,

$$\begin{aligned} z_{v,i} &= u_i^\top \widehat{x}_- / \sqrt{\kappa} + b_i, \qquad z_{k,i} = k_i^\top y_+, \\ \widehat{x}_+ &= \widehat{x}_- + \sqrt{\kappa} u_i (-\beta_i z_{v,i} + w_i z_{k,i}), \qquad y_- = y_+ - k_i (g_i z_{v,i} + \alpha_i z_{k,i}). \end{aligned}$$

Initial and terminal conditions are $\widehat{x}_0 = M_0 y_0$ and $y_{\mathrm{end}} = 0$. Every intervening right contraction $Q$ has $\widehat{x}_+ = \widehat{x}_-$, $y_- = Q y_+$, with no forcing. Forward summation gives the strict-lower value Gram matrix; backward summation gives the ordered products above, hence $\mathcal{F} z = (b, 0)^\top$.

**Defect propagation.** Set

$$L_i = I - \beta_i u_i u_i^\top, \quad R_i = I - \alpha_i k_i k_i^\top, \quad W_i = \sqrt{\kappa} w_i u_i k_i^\top, \quad G_i = -g_i k_i u_i^\top / \sqrt{\kappa}.$$

The scattering matrix $\left(\begin{smallmatrix} L_i & W_i \\ G_i & R_i \end{smallmatrix}\right)$ is orthogonal. On nonzero active directions it is $\left(\begin{smallmatrix} r_i & \tau_i \\ -\tau_i & r_i \end{smallmatrix}\right)$, where $r_i = c_i^{-1}$, $\tau_i = \sqrt{\kappa}\|u_i\|\|k_i\|/c_i$ and $r_i^2 + \tau_i^2 = 1$; if either vector is zero, it is identity. The forced equations are therefore

$$\begin{aligned} \widehat{x}_+ &= L_i \widehat{x}_- + W_i y_+ + f_i, & y_- &= G_i \widehat{x}_- + R_i y_+ + h_i, \\ f_i &= -\sqrt{\kappa} \beta_i u_i b_i, & h_i &= -g_i k_i b_i. \end{aligned}$$

Coefficient bounds give $\|f_i\| \le \kappa K_* (t_*/2) |b_i|$ and $\|h_i\| \le \kappa K_* |b_i|$.

Along the actual unforced states write $\widehat{x} = My + p$. At a write with incoming $M$, put

$$\Xi_i = (I - G_i M)^{-1}, \quad \mathsf{A}_i = L_i (I - M G_i)^{-1}, \quad \mathsf{D}_i = \Xi_i R_i.$$

Here $\|\Xi_i\| \le a_*$ because $\|G_i\| \le 1$ and $\|M\| \le s$. Direct substitution yields

$$p_+ = \mathsf{A}_i p_- + L_i M \Xi_i h_i + f_i, \qquad y_- = \mathsf{D}_i y_+ + \Xi_i G_i p_- + \Xi_i h_i.$$

Orthogonality of the scatterer gives the classical defect identities

$$\begin{aligned} \mathsf{A}_i (I - M M^\top) \mathsf{A}_i^\top &= I - M_+ M_+^\top, \\ \mathsf{D}_i^\top (I - M^\top M) \mathsf{D}_i &= I - M_+^\top M_+. \end{aligned}$$

They follow by substituting $M_+ = W_i + L_i M \Xi_i R_i$ and the orthogonality block identities. At a right contraction, $M_+ = MQ$, $p_+ = p_-$, $\mathsf{A} = I$, $\mathsf{D} = Q$. The equalities become $\preceq$; their positive defects are $M(I - QQ^\top)M^\top$ and $I - Q^\top Q$. Since all state defects lie between $(1 - s^2)I$ and $I$, telescoping bounds every forward or backward propagator by $\nu$, regardless of its length.

**Inverse estimate.** The forward injection is bounded by

$$\|L_i M \Xi_i h_i + f_i\| \le \|f_i\| + sa_*\|h_i\| \le \kappa K_*(t_*/2 + sa_*)|b_i|.$$

Starting from $p_0 = 0$, Cauchy–Schwarz over $n$ writes gives $P := \max\|p\| \le \nu\kappa K_*(t_*/2 + sa_*)\sqrt{n}\|b\|$. Backward injection is at most $a_*(P + \|h_i\|)$, so $y_{\mathrm{end}} = 0$ gives

$$Y := \max\|y\| \le \nu a_*[nP + \kappa K_*\sqrt{n}\|b\|] \le \nu a_*\kappa K_*\sqrt{n}[n\nu(t_*/2 + sa_*) + 1]\|b\|.$$

Finally $\|\widehat{x}\| \le sY + P$ gives the value-port bound $\|z_v\| \le (U_*/\sqrt{\kappa})\sqrt{n}(sY + P) + \|b\| \le C_n\|b\|$. The recurrences construct a solution for every value forcing. For homogeneous forcing they uniquely give $p = 0$, then $y = \widehat{x} = z = 0$, so the square matrix $\mathcal{F}$ has trivial kernel. Hence it and $H$ are invertible, and $z_v = H^{-1}b$ proves the inverse bound. Zero ports are included. At $u_i = 0$, for example, $z_{v,i} = b_i$ but $h_i = -\kappa k_i b_i$ need not vanish. Keeping this forcing, rather than deleting that identity write, retains the corresponding column.

**Norm and condition number.** The ordered products and $\|S_0\| \le s/\sqrt{\kappa}$ imply $|G_{ij}| \le U_*^2$ and $|\Omega_{ij}| \le U_* K_* s/\sqrt{\kappa}$. Using $\beta_i \le \kappa K_*^2/2$, $\alpha_i \le \kappa U_*^2/2$, $w_i \le 1$, $g_i \le \kappa$ gives

$$\|A_v\| \le 1 + n(t_*^2/2 + st_*), \qquad \|B_v\| \le nU_*^2(1 + st_*/2).$$

The finite triangular expansion of $Z_d = K_s^{-1}C$ groups into

$$(Z_d)_{ij} = g_j k_i^\top Q_{i+1}R_{i+1}\cdots Q_{j-1}R_{j-1}Q_j k_j \quad (i < j),$$

and zero otherwise. Every $Q_i, R_i$ contracts, so $|(Z_d)_{ij}| \le \kappa K_*^2$ and $\|Z_d\| \le n\kappa K_*^2$. Thus $\|H\| \le \|A_v\| + \|B_v\|\|Z_d\| \le B_n$, which proves $\mathrm{cond}_2(H) \le B_n C_n$. At $\kappa = 0$, $\beta = \alpha = g = 0$ gives $H = I$ directly.

**Why the margin is needed.** Set $V = K = 1$, $\kappa = 1$, $S_0 = 0$, $D = 1$, $e = 0$, $k_i = 1$, and $n = 2N$. Take $u_i = \sinh\eta$ for $i \le N$, then $u_i = -\sinh\eta$, with fixed $\eta > 0$. The net transfer is identity but its midpoint state is $\tanh(N\eta)$. Let $r = \operatorname{sech}\eta$. For terminal key one, the triangular key solution is $(z_0)_i = r^{n-i}$ and the value right-hand side obeys $|(-B_v z_0)_i| \le \sinh^2\eta\, r^2/(1-r) = 1 + r$. The midpoint transfer coordinate is $x_N = \sinh(N\eta)$, giving $z_{v,N+1} = -\sinh\eta\sinh(N\eta)$. Hence

$$\|H^{-1}\| \ge \frac{\sinh\eta\sinh(N\eta)}{(1 + \operatorname{sech}\eta)\sqrt{2N}}.$$

Thus bounded controls without a uniform interior margin allow exponential inverse growth; these controls have no strict decay.

# D Supplementary experiments

## D.1 Matched state tracking

All three models have width 256, two layers, four heads and $K = V = 64$, giving 32,768 recurrent matrix-state scalars. Curved and Tangent each have 7,733,471 parameters; Native has one fewer. Three independent initializations are shared across methods, together with the complete-game batch order. The shared curvature parameter starts at $a = 1/8$, and the proposed memories use the generic $n = 32$ chunk solver.

Training batches contain 16 games with right padding and no length cap. AdamW uses a constant learning rate $10^{-3}$, momentum coefficients $(0.9, 0.95)$, $\epsilon = 10^{-8}$ and gradient-norm clipping at one. Weight decay is $0.01$ for matrix parameters, preserving the native layer's exclusions, and zero for vectors and scalars. The frontend uses BF16 autocast; master parameters and the proposed memory core use FP32. Training ends at the first complete batch reaching the prescribed number of nonpadding moves. All 75 labels, including START, enter the loss; reported prediction accuracy excludes START.

The 128M study uses 989,847 March training games and 10,153 validation games. May, June and Random each contain 10,000 evaluation games, with 689,286, 689,610 and 3,306,304 postmove states, respectively. May and June use Lichess game records. Random samples each move uniformly

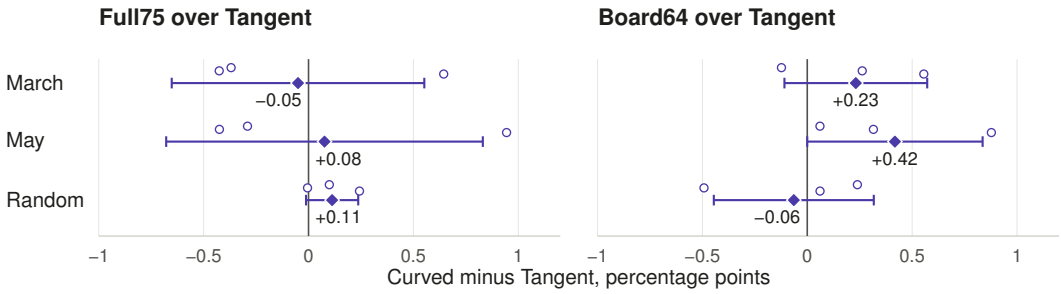


Figure 6: **The independent 32M paired study.** Differences of Curved from Tangent over three paired runs trained on 226,524 games, with mean and one sample SD.

from the legal moves until a terminal position or claimable draw, then retains games of at least 20 plies. Games are not truncated. Identity and full-trajectory checks found no training overlap. June was selected before its evaluation outcomes. The cohorts' length and move distributions vary jointly.

Means and sample standard deviations describe the three training runs; contrasts are computed within matched runs before aggregation. The independent 32M study of Curved and Tangent uses 226,524 training and 2,403 validation games, with three paired initializations and the same model shell. Figure 6 reports it separately because its training corpus differs from the 128M study.

### D.2 Supplementary mechanism analysis

The trajectory analysis uses the first 128 complete March validation games for development and the first 256 complete June games for confirmation, across three runs of the curved and tangent families. Component interventions use the curved models on those June games. Complete-network swaps and endpoint comparisons use the separate 10,000-game June cohort.

The actual frontend supplies all controls. Local comparisons restart at each captured chunk boundary and advance the original baseline within the chunk. Persistent FP32 curved memory $C_t$ and tangent trajectory $\widetilde{C}_t$ instead share an initial state and continue across chunks, without changing the original network output. With $T_t(X) = L_t X R_t + w_t u_t k_t^\top$,

$$\Delta_t = F_t(C_{t-1}A_t) - T_t(C_{t-1}A_t), \qquad E_t = C_t - \widetilde{C}_t = L_t E_{t-1} A_t R_t + \Delta_t.$$

Sequential-versus-chunk reads and states, together with separately materialized BF16, normalized and projected outputs, provide finite-precision replay references rather than strict error bounds.

For $P_v = vv^\top/\|v\|^2$ when $v \neq 0$ and $P_0 = 0$, define

$$\Delta_t^\perp = (I - P_{u_t})\Delta_t(I - P_{k_t}), \quad E_t^\perp = L_t E_{t-1}^\perp A_t R_t + \Delta_t^\perp, \quad E_t^\parallel = E_t - E_t^\perp.$$

The insertion components $\Delta_t^\perp$ and $\Delta_t - \Delta_t^\perp$ are Frobenius-orthogonal. Their transported histories need not remain orthogonal to each other or to the current write directions. Signed cross terms are therefore retained.

Each intervention replaces one layer's normalization input by $C_t q_t$, $\widetilde{C}_t q_t$, $(C_t - E_t^\perp)q_t$, or $(C_t - E_t^\parallel)q_t$. The actual gate, BF16 materialization, learned normalization and projection are followed by the residual connections, MLP, remaining layers and task head. Later layers recompute controls after an earlier-layer intervention; component histories retain the original curved controls. These are separate read interventions. Each full branch is normalized independently, so its downstream effect is not an additive energy contribution. Predictions are paired with the contemporaneous original forward; the $C_t q_t$ branch measures replay's prediction effect.

For differences $d_i$ and corresponding references $r_i$, relative RMS is

$$\mathrm{rRMS}(d; r) = \left(\frac{\sum_i \|d_i\|^2}{\sum_i \|r_i\|^2}\right)^{1/2}.$$

Table 4: Accuracy change after removing one remainder history from one layer's read, on the first 256 June games. Intervention minus original forward in pp, mean and sample SD over three curved checkpoints. Changed and unchanged refer to piece labels.

| Layer | Removed | Full75 | Board64 | Counter | Changed | Unchanged |
|---|---|---|---|---|---|---|
| 0 | $E$ | $-1.88 \pm 1.50$ | $-1.20 \pm 0.63$ | $-0.92 \pm 1.67$ | $-0.04 \pm 0.05$ | $-0.02 \pm 0.01$ |
| 0 | $E^{\perp}$ | $-1.57 \pm 1.67$ | $-0.79 \pm 0.62$ | $-1.13 \pm 2.05$ | $-0.00 \pm 0.04$ | $-0.02 \pm 0.01$ |
| 0 | $E^{\parallel}$ | $-0.21 \pm 0.25$ | $-0.22 \pm 0.21$ | $+0.03 \pm 0.11$ | $-0.01 \pm 0.03$ | $-0.00 \pm 0.01$ |
| 1 | $E$ | $-62.32 \pm 16.38$ | $-16.08 \pm 2.88$ | $-71.19 \pm 20.33$ | $-2.14 \pm 0.39$ | $-0.36 \pm 0.07$ |
| 1 | $E^{\perp}$ | $-51.53 \pm 20.85$ | $-14.88 \pm 2.95$ | $-58.71 \pm 21.51$ | $-1.88 \pm 0.38$ | $-0.32 \pm 0.07$ |
| 1 | $E^{\parallel}$ | $-34.90 \pm 8.82$ | $-1.13 \pm 0.26$ | $-45.18 \pm 10.51$ | $-0.15 \pm 0.05$ | $-0.02 \pm 0.00$ |

Table 5: June field accuracy in percent, mean over three runs on 10,000 games. Subscripts give the training and the evaluation write.

| Field | $C_C$ | $C_T$ | $T_T$ | $T_C$ |
|---|---|---|---|---|
| Pieces | 99.726 | 99.253 | 99.700 | 99.587 |
| Side to move | 99.986 | 90.237 | 99.980 | 99.981 |
| Castling | 99.991 | 99.982 | 99.992 | 99.988 |
| En passant file | 99.984 | 99.884 | 99.987 | 99.969 |
| En passant rank | 99.996 | 99.681 | 99.996 | 99.977 |
| Half-move clock | 96.714 | 91.528 | 96.932 | 94.559 |
| Full-move counter | 98.102 | 28.903 | 98.039 | 89.100 |
| Full-move digit fields | 99.051 | 64.452 | 99.020 | 94.550 |

Sums pool games, positions and heads within each model and layer before taking the ratio; summaries give mean and sample SD across models. Zero denominators remain undefined. Local raw summaries use the corresponding replayed baseline. Persistent read summaries compare the original curved read with the tangent trajectory and use that original read as reference. Position bins share this actual curved-read reference; component readout comparisons use the actual output at each stage. Prediction metrics exclude START. Energy summaries include it, while position-bin curves exclude it. Bins 1–16, 17–32, 33–64, 65–128 and 129–256 contain 4,096, 3,961, 5,758, 3,029 and 114 tokens from 256, 256, 232, 119 and 8 games, respectively; no later positions occur.

The fixed example uses the first training run, the first June game reaching token31, layer1/head0, with START=0. Its orthonormal active bases are

$$U_{\mathrm{act}} = [\widehat{u}, \widehat{(I - P_u)LXk}], \qquad K_{\mathrm{act}} = [\widehat{k}, \widehat{(I - P_k)R^{\top}X^{\top}u}],$$

where $\widehat{v} = v/\|v\|$ for nonzero $v$, and zero otherwise. The plotted query is $K_{\mathrm{act}}^{\top} q$ and the correction is $U_{\mathrm{act}}^{\top} \Delta K_{\mathrm{act}}$. Component reconstruction residuals are below $1.4 \times 10^{-8}$ in absolute Frobenius norm. This preselected snapshot and the pooled three-model read-effect summary answer different questions.

The complete counters are two base-256 digits. On both the 256-game and the 10,000-game June sets, the high digit of each counter is correct for every state in every branch, so complete counter accuracy equals low-digit accuracy and the average over the two digit fields equals $(100{+}\text{counter})/2$. Table 4 also splits piece labels into those that change at the current move and those that do not. The two pools differ in size by a factor of about 30 and in baseline error, so their losses are reported side by side and not compared as rates of one mechanism.

### D.3 Computational cost and numerical accuracy

Table 3 measures a complete forward pass, scalar task loss and backward pass using the first matched 128M run of each method. The common 16-game batch contains 1,030 moves and pads to 121 positions. Two warmups precede five synchronized repeats on an RTX PRO 6000 Blackwell GPU. The median excludes loading, gradient clearing, clipping and optimizer updates. Incremental memory is peak allocated memory above the pre-forward baseline. The frontend runs in BF16; master parameters, loss logits and the proposed memory core use FP32.

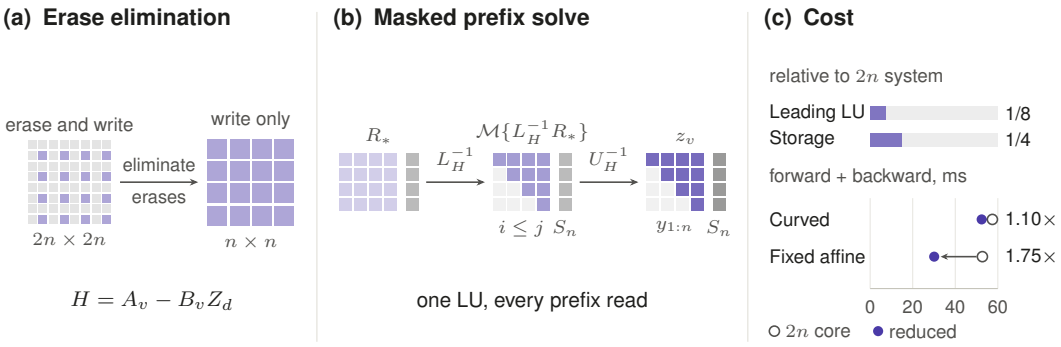


Figure 7: **Exact computation after erase elimination.** One unknown remains per token, and the mask between the two triangular solves returns every prefix read and the final state. (c) Algebraic cost ratios and the measured forward and backward pass.

The specialized tangent replaces generic LU by triangular solves, preserving derivatives through $\beta, \alpha, w$ and curvature. On this batch, its BF16 task logits and FP32 scalar loss exactly match the generic tangent's; the curvature-gradient relative RMS discrepancy is $4.80 \times 10^{-6}$. Its whole-model speedup is $1.18\times$. Reported training endpoints use the generic core. Figure 7 compares the formulation with separate erase and write unknowns and the reduced one. Eliminating erase unknowns reduces leading dense LU work eightfold and interval storage fourfold. A separate frozen-model benchmark after 32M moves gives whole-model speedups of $1.75\times$ for the fixed-affine boundary and $1.10\times$ for Curved.

The checks in Table 2 use saved controls of a trained language-model layer on 512 FineWeb tokens, a nonzero FP64 initial state and chunk lengths 32 and 64. Four cases cover the original controls, signed erasure, signed erasure with exact zero decays, and the affine boundary $a = 0$, all at $a = 1$ otherwise. Gradients cover the initial state, the value, key, query, erase and decay inputs, and the curvature parameters $a$ and $\kappa$. Relative maximum divides the largest absolute difference by the largest absolute reference entry, maximized over cases. The reduced and the two-event formulations agree with each other more closely than either agrees with the direct recurrence.

# E Supplementary Task Comparisons

## E.1 State Tracking with Alternative Cells

A 32M-move CWM comparison evaluates curved memory, its zero-curvature boundary, native GDN2 and KDN in a common shell. Across three runs, all six cohort/comparator mean contrasts favor curved memory, although the May contrast with native GDN2 changes sign across runs. The boundary changes both the empty-state write and its Jacobian. KDN additionally stores 512 precision scalars alongside 32,768 content scalars. The closer curved/tangent comparison gives small, mixed differences in Figure 6. Its corpus differs from the 128M study.

## E.2 Recall, Accumulation, Overwrite and Deletion

Synthetic tasks test exact recall after accumulation, overwrite and deletion. Each event uses three tokens. Exact-query accuracy requires every label requested at a query to be correct; dynamic streams score both interleaved and final queries. With eight entities, training/testing uses 12/24 accumulation attributes, 4/16 overwrites or 24/48 dynamic events per entity. Deletion chooses a uniformly random live attribute; mixed streams combine deletion and overwrite. Length and load change jointly except in overwrite, which fixes entity count. Variants increase entities from eight to sixteen.

The compared geometric cells use unit $u, k$ and the time-one simultaneous flow

$$\dot{M} = -\nu u u^\top M - \nu M k k^\top + b_0 u k^\top - c_0 M k u^\top M.$$

Table 6: Synthetic exact-query accuracy at increased load or history. Mean ± sample SD over three runs; stars mark two-run estimates. Mixed combines deletion and overwrite. The dash indicates an unreported result.

| Cell | Accumulation | Overwrite | Deletion | Mixed |
|---|---|---|---|---|
| Riccati, coupled $\nu$ | $.980 \pm .011$ | — | $.996 \pm .003$ | $.981 \pm .011^*$ |
| Bounded affine | $.963 \pm .034$ | $1.000 \pm .000$ | $.976 \pm .010$ | $.951 \pm .065$ |
| Riccati, independent $\nu$ | $.869 \pm .106$ | $1.000 \pm .000$ | $.998 \pm .002$ | $.880 \pm .099$ |
| DeltaNet | $.055 \pm .064$ | $1.000 \pm .000$ | $.933 \pm .055$ | $.980 \pm .015$ | |
| Gated DeltaNet | $.000 \pm .000$ | $1.000 \pm .000$ | $.848 \pm .046$ | $.879 \pm .096$ |
| Gated DeltaNet-2 | $.496 \pm .499^*$ | $1.000 \pm .000$ | $.923 \pm .041$ | $.959 \pm .029$ |

Table 7: Controller adaptation of the pretrained hybrid. WikiText reports word perplexity; LAMBADA and retrieval report accuracy in percent.

| Memory | WikiText PPL | LAMBADA | Retrieval |
|---|---|---|---|
| Native | 8.6739 | 72.56 | 59.173 |
| Flat | 8.6723 | 72.48 | 59.194 |
| Curved | 8.6746 | 72.58 | 59.181 |

Bounded affine sets $c_0 = 0$ and $\nu = \sqrt{(b_0/2)^2 + \mathrm{softplus}(r)^2}$. The Riccati variants set $c_0 = b_0$, with either this coupled $\nu$ or independently learned $\nu = \mathrm{softplus}(r)$. The condition $\nu \geq |b_0 - c_0|/2$ preserves the unit ball. These flows differ from the split recurrence, and dissipation means deletion is learned without exact boost inversion. Accumulation and deletion depend on the cell, while overwrite saturates (Table 6). The comparisons include both update and normalization differences.

### E.3 Language modeling across context lengths

A positive-erase split with width 256 and six layers trains on 32.8M FineWeb-Edu tokens. Contexts from 512 to 4096 tokens score the same 14,336 FineWeb targets and 6,400 PG19 targets. Fixed curvature ($\kappa = 1$) performs worse than affine at every length; all primary learned-minus-affine paired-resampling intervals include zero. The learned shared curvature ends at .05843. These intervals describe document variation conditional on one training run.

Target-centered cells that preserve an association fiber use a different affine comparator. After 536.9M tokens in a width-512, twelve-layer model, curved/affine WikiText-2 word perplexities are $104.84/103.00$, versus $90.67/90.13$ for native GDN2/KDN. Both target-centered cells also trail the native models on LAMBADA and sampled PG19 full-prefix targets. The PG19 measurement covers 128 targets per book and therefore measures sampled prediction quality rather than full-book perplexity.

### E.4 Pretrained adaptation and retrieval

Qwen3.8-27B combines 48 GDN and 16 attention layers. Controller adaptation trains 23.6M existing parameters on identical 1.05M-token FineWeb-Edu inputs; curved adds one scalar. Flat uses the same recurrence as Native in exact arithmetic, evaluated by the custom FP32 chunk solver. Table 7 shows near parity on WikiText-2, LAMBADA and six retrieval tasks. Retrieval averages SWDE, SQuAD, FDA, TriviaQA, NQ and DROP using identical answer-aware crops and case-insensitive substring scoring. Curved improves two tasks, decreases two and ties two relative to flat memory.

Adapting 2.54B query, key, value and controller parameters places curved first on WikiText-2 perplexity and LAMBADA perplexity/accuracy, but its LAMBADA accuracy exceeds flat by one example. Every arm trails controller-only adaptation. These single-run hybrid results show that trainable subspace and task matter when translating a memory operation into a quality gain.