OSCILLATORY DELTA NETWORKS: PHASE-ADDRESSED ASSOCIATIVE MEMORY

Anonymous authors
Paper under double-blind review

ABSTRACT

Linear attention and its delta-rule variants, such as Gated DeltaNet, act as implicit associative memories with a fixed-size matrix state. Such a memory reaches full capacity only with orthogonal keys, yet these models have no explicit inductive bias toward them. Our analysis shows that trained DeltaNets generate highly correlated keys and therefore overwrite one another. To overcome such interference, we introduce *Oscillatory Delta Networks* (ODN), with an inductive bias decorrelate keys by transporting the memory along volume-preserving orbits before each delta-rule write. For rotation based ODNs (ODN-SO(2)) one input-dependent clock per head advances a spectrum of oscillators. In this framework each key-value association is addressed by both its key and its internal write time. We prove that under a suitable spectrum and clock, this oscillation-based content addressing attains the memory's theoretical capacity limit even when all writes share one key, whereas static content addressing retains only the most recent value. Extending memory transports to special linear transformations (ODN-SL(2)) additionally learns each orbit's geometry under the same clock. In language modeling, ODN-SO(2) and ODN-SL(2) outperforms Gated DeltaNet, Selective Rope, and Kimi Delta Attention (KDA). Experiments show this route to be also more effective and parameter-efficient than the channel-wise gating (KDA, Selective Rope). Our approach, inspired by oscillatory phase-dependent short-term memory encoding in biological brain circuits, reduces the overlap between addresses of repeated tokens and enables efficient modeling of correlated long-range interactions.

1 INTRODUCTION

Linear attention and selective state-space models process long sequences with a fixed-size recurrent state (Katharopoulos et al., 2020; Gu & Dao, 2024), instead of the key–value cache that grows with context in softmax attention (Vaswani et al., 2017). Delta-rule models such as DeltaNet, Gated DeltaNet, and Kimi Delta Attention (KDA) update this state by writing key - value associations corrected by the memory's already existing predictions, and employing adaptive forgetting to control memory retention(Schlag et al., 2021; Yang et al., 2025; Kimi Team, 2025). Their central limitation is interference between associations. In the classical correlation-matrix view, arbitrary associations can be reliably recalled without cross-talk only when their key addresses show low pairwise key correlations (Kohonen, 1972; Schlag et al., 2021) (Kohonen, 1972; Schlag et al., 2021). Existing delta-rule architectures (Schlag et al., 2021; Yang et al., 2025; Kimi Team, 2025) do not explicitly decorrelate their learned write keys and therefore provide no structural guarantee that distinct associations receive distinct addresses. When two keys overlap, a later update along one key necessarily modifies the memory read through the other.

In line with observations that keys in standard transformer based large language models (LLMs) occupy substantially lower-dimensional subspaces than their nominal dimension
Singhania et al. (2024), our analysis finds that write keys in DeltaNet-family models also are correlated and occupy low effective dimensional subspaces. This structure can accelerate learning, improve generalization across similar inputs, and simplify query - key alignment. However, it reduces the capacity for independent associations, increases crosstalk between incompatible writes, and leaves much of the nominal state weakly addressable. Standard DeltaNet formulations impose no explicit constraint to control this key geometry, motivating our model.

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.

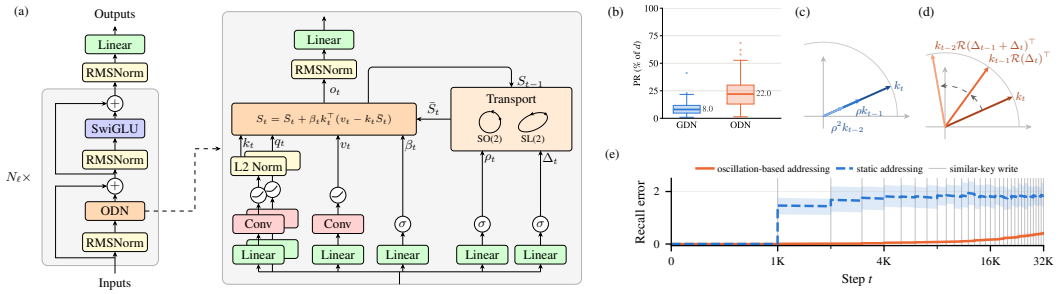

Figure 1: Oscillatory Delta Networks and rotation in associative memory. **(a)** ODN architecture: transport maps S_{t-1} to $\bar{S}_t$ before the delta update. ODN-SO(2) uses circular orbits; ODN-SL(2) learns volume preserving orbit geometry (Equations 2 and 3). **(b)** Intrinsic dimensionality of the subspace spanned by the set of 4k address keys obtained for a pretrained 340M Gated DeltaNet (GDN) (left) and ODN-SO(2) (right) using 128 documents of WikiText-2. Subspace dimensionality measured by the participation ratio (PR) as a percentage of the key dimension $d = 128$. over 192 layer–heads, with medians labeled (Table 1). Low intrinsic dimensionality indicates high correlated keys. **(c)** Static content addressing with scalar retention keeps the previous keys k_{t-1} and k_{t-2} aligned with the current key k_t, so the new write operations overwrite the old associations. **(d)** Oscillation-based content addressing rotates the state between writes, carrying the past addresses k_{t-1} and k_{t-2} to different directions from k_t according to their elapsed internal times. This separation reduces overwriting. (e) The recall error $\|o_t - v_0\|_2^2$ at time t for static and oscillation-based content addressing. Keys similar to the target key k_0 recur every 1K steps (gray). Oscillatory addressing preserves more of the target than in the static case.

Our idea is to transport the state memory before each new write (Figure 1). Specifically the ODN first applies an input-dependent volume-preserving transformation to the recurrent state, after which the delta rule writes the incoming key–value pair. This conserves the magnitude of the stored information while changing its alignment with the incoming key. Thus, an earlier association can be moved away from the direction targeted by a repeated key, reducing direct overwriting without erasing old associations.

Our transformations are block-wise planar rotations (**ODN-SO(2)**) or special linear transformations (**ODN-SL(2)**). ODN-SO(2) applies rotations using a spectrum of frequencies distributed over the blocks and is driven by a input-dependent internal time advance. ODN-SL(2) generalizes this to equivolumetric transformations. Inspired by how different short-term memories in biological brains are best encoded at different phases of oscillatory brain dynamics (Lisman & Idiart, 1995; Siegel et al., 2009; Lisman & Jensen, 2013), we term this memory transport *oscillation-based content addressing* . The controlled continuous-write experiment in Figure 1(e) connects oscillation-based address separation to improved target preservation under similar-key interference (Appendix H.2 gives the details).

The mechanism is a recurrent counterpart of rotary position embeddings (RoPE) (Su et al., 2021): one shared displacement is represented across a spectrum of frequencies (Barbero et al., 2025), but ODN measures content-dependent internal time rather than token position.

Our contributions are:

- **Interference from correlated keys.** Delta-rule models are implicit associative memories that need orthogonal keys for full capacity, yet we show that trained DeltaNet-family models generate highly correlated keys, so the delta rule erases non-selectively along a few shared directions and leaves many directions of the state unused (Sections 4.1 and 5.1).
- **Oscillation-addressed associative memory.** ODN-SO(2) transports the memory along oscillatory orbits before each delta-rule write, so each association is addressed by its key and its internal write time. We prove that this addressing can attain the memory's theoretical capacity limit even when all writes share one key (Sections 3 and 4).

- **Added expressivity in the orbit geometry.** ODN-SL(2) extends the transport to special linear maps that learn each orbit's geometry under the same clock, which proves both more effective and more parameter-efficient than the commonly used channel-wise gating (Sections 5.4 and 5.5).
- **Empirical gains.** At matched state size, ODN outperforms Gated DeltaNet, KDA, and Selective RoPE on associative recall, state tracking, and language modeling (Section 5).

2 Related Work

Delta-rule and selective memory. Fast weight programmers and linear attention store associations in temporary weight matrices (Schmidhuber, 1992; Katharopoulos et al., 2020). DeltaNet adds error-correcting writes and efficient parallel training (Schlag et al., 2021; Yang et al., 2024); Gated DeltaNet and KDA add adaptive retention (Yang et al., 2025; Kimi Team, 2025), and DeltaProduct enlarges the transition family with products of Householder maps (Siems et al., 2025). Selective SSMs likewise make state evolution input-dependent (Gu & Dao, 2024; Dao & Gu, 2024). ODN retains the delta rule while introducing an input-dependent memory transport that precedes its residual computation.

Oscillatory phase as a memory index in the brain. ODN is conceptually motivated by neurobiological evidence that oscillatory phase and evolving temporal context can separate otherwise overlapping memories. The theta–gamma model proposes that multiple memory items are held in one neuronal network by occupying them at different phases within a theta oscillation cycle (Lisman & Idiart, 1995; Lisman & Jensen, 2013), consistent with phase-dependent coding of item order in monkey prefrontal cortex and human episodic memory (Siegel et al., 2009; Heusser et al., 2016). Over longer intervals, CA1 ensembles evoked by the same context progressively diverge over hours to days while CA3 remains stable (Mankin et al., 2012), and episodes close in time share a CA1 timestamp while remote ones, even in the same context, do not (Rubin et al., 2015). The temporal context model formalizes such an index: each item is bound to a slowly drifting context vector that later serves as its retrieval cue (Howard & Kahana, 2002). ODN uses an internal time driven state transport to update matrix memories and addresses.

Oscillations and phase in sequence models. Complex recurrences, LinOSS, and learned dissipation establish oscillation and damping as tools for sequence modeling (Orvieto et al., 2023; Rusch & Rus, 2025; Boyer et al., 2025); Mamba-3 combines complex dynamics with selective state updates (Lahoti et al., 2026). Phase binding predates these models in holographic and recurrent associative memories (Plate, 1995; Danihelka et al., 2016). Closest to our transport, RetNet applies a fixed rotation and decay to a linear-attention state (Sun et al., 2023), RoPE uses a shared positional clock (Su et al., 2021), and Selective RoPE learns frequency-wise phase drives (Movahedi et al., 2026). ODN-SO(2) instead applies an input-dependent common clock and persistent spectrum to a delta-rule memory; ODN-SL(2) adds learned orbit geometry. Extended comparisons appear in Appendix B.

3 Oscillatory Delta Networks

3.1 From Gated DeltaNet to Transported Delta Writes

Linear attention admits a recurrent form in which a fixed-size state matrix stores key–value associations and is queried at each token (Katharopoulos et al., 2020). Gated DeltaNet augments this recurrent view with adaptive retention and error-correcting delta-rule writes (Yang et al., 2025). More precisely, in a single head at token index t , learned projections of the input representation x_t produce a key $k_t \in \mathbb{R}^{1\times d}$, a query $q_t \in \mathbb{R}^{1\times d}$, and a value $v_t \in \mathbb{R}^{1\times d_v}$. These modify a memory state matrix $S_t \in \mathbb{R}^{d\times d_v}$ ($S_0 = 0$) that at each step is scaled by an adaptive scalar retention gate $\rho_t \in (0, 1)$ and updated by writing the residual of the current key–value association controlled by the write gate $\beta_t \in (0, 1)$ produce the output o_t according to:

$$S_t = \rho_t S_{t-1} + \beta_t k_t^\top (v_t - k_t S_{t-1}), \qquad o_t = q_t S_t. \tag{1}$$

ODN uses the same variables and keeps the projections, prediction based residual delta write, and readout unchanged. It replaces only the scalar transition $\rho_t I$ before the write in Gated DeltaNet by a structured linear map $\rho_t T_t$:

$$\bar{S}_t = \rho_t T_t S_{t-1}, \quad S_t = \bar{S}_t + \beta_t k_t^\top \left(v_t - k_t \bar{S}_t\right). \tag{2}$$

3.2 A Shared Clock and Learned Orbits

For ODN, let $d = 2m$ and group the d state rows into m consecutive pairs, called oscillators, with rows $2r$ and $2r+1$ forming oscillator r. Each planar oscillator follows an orbit defined by a 2×2 generator matrix A_r. One input-dependent step Δ_t per head advances all oscillators along their respective orbits:

$$T_t = \bigoplus_{r=1}^{m} T_{t,r}, \qquad T_{t,r} = \kappa_{t,r} \exp(\Delta_t A_r). \tag{3}$$

Here we use uniform $\kappa_{t,r} = 1$ for ODN-SO(2); ODN-SL(2) uses the hyperbolic guard defined below. This separates the orbit geometry, learned across training, from the internal time that drives transformations and is selected at each token.

Rotational orbits: ODN-SO(2). ODN-SO(2) uses a skew-symmetric generator

$$A_r = \omega_r J, \qquad J = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}, \qquad T_{t,r} = \exp(\Delta_t \omega_r J), \tag{4}$$

for each oscillator r where $\omega_r > 0$ is its learned rotation frequency. Thus oscillator r advances by phase $\theta_t = \omega_r \Delta_t$, and the frequency spectrum ω jointly encodes the shared scalar internal time.

Equivolumetric orbits: ODN-SL(2). To allow training to reshape the orbits but keep the memory volume preserved, we learn

$$A_r = \omega_r \begin{bmatrix} \phi_r & -1+\psi_r \\ 1+\psi_r & -\phi_r \end{bmatrix}, \qquad \delta_r = 1 - \phi_r^2 - \psi_r^2. \tag{5}$$

Its zero trace ensures $\det \exp(\Delta_t A_r) = 1$. The sign of δ_r selects elliptic (deformed rotation), parabolic (shear), or hyperbolic (expanding/contracting) dynamics. Initialization at $\phi_r = \psi_r = 0$ recovers SO(2). The extension keeps the common internal time: every oscillator obeys $\exp(\Delta A_r)\exp(\Delta' A_r) = \exp((\Delta + \Delta')A_r)$ for its learned generator.

Hyperbolic guard. For SL(2), a hyperbolic transformation ($\delta_r < 0$) has an eigenmode that grows at rate $\omega_r\sqrt{-\delta_r}$ per unit of internal time; we cancel this growth with an equal decay driven by the same clock by setting

$$\kappa_{t,r} = e^{-\omega_r^+ \Delta_t}, \qquad \omega_r^+ = \omega_r \sqrt{\max(-\delta_r, 0)}, \tag{6}$$

which leaves elliptic and parabolic regimes untouched and acts as a hyperbolic-specific forget gate. Explicit transport matrices, exact exponentials, and stability details are provided in Appendix C.

Frequency initialization. We initialize each oscillator frequency by sampling uniformly in log space,

$$\log \omega_r \sim \mathcal{U}(\log \omega_{\min}, \log \omega_{\max}), \qquad r = 1, \ldots, m, \tag{7}$$

where $[\omega_{\min}, \omega_{\max}]$ is a frequency-range hyperparameter. The frequencies remain learnable during training.

Chunkwise computation and cost. The WY/UT construction of Gated DeltaNet and KDA (Yang et al., 2025; Kimi Team, 2025) parallelizes both variants within chunks; the only change is that scalar interval retention factors become 2×2 transport blocks, which are never expanded into a dense $d \times d$ matrix. For sequence length L and chunk width b, forward work per head is $O(Ldd_v + Lb(d + d_v) + Lb^2)$, the same order as Gated DeltaNet, and decoding carries the same $d \times d_v$ state. Appendix G gives the algorithm.

4 Similar Keys and Oscillation-Based Content Addressing

A delta write erases earlier associations in proportion to their address overlap, so the overlap of similar keys decides what the memory keeps. This section shows that this static content addressing forces similar keys to overwrite one another, while oscillation-based addressing separates them

through a temporal kernel. The geometric frequency spectrum which ODN inherits from RoPE makes this kernel an inductive bias: content repeated close in internal time is corrected, whereas content repeated far apart is retained. We analyze ODN-SO(2) with unit retention and read each association with a query matched to its key address. Full proofs are given in Appendix E.

4.1 Similar Keys Overwrite their Memories in Static Addressing

Let association i have unit key $\hat{k}_i$, and let G_{ij} be the overlap of the addresses under which associations i and j are stored. A delta write at j changes the stored contribution of an earlier association i in proportion to G_{ji} (Appendix E.1). For static content addresses, $G_{ij} = \hat{k}_i \hat{k}_j^\top$ is fixed by the keys alone, regardless of when the writes occur. If all pairwise key overlaps are at least c_0, delta-rule recall of the older of two associations therefore has mean squared error at least c_0^2, and for identical keys only the most recent value survives (Proposition 1). Neither bound involves the state dimension d. Scalar retention cannot help either: it rescales stored associations without changing their address directions, so every G_{ij} is unchanged. It only attenuates older associations more than recent ones. Similar keys are common in a pretrained Gated DeltaNet (Figure 1b) and repeated occurrences of one token have a mean squared key overlap about fifty times the value $1/d$ expected for random keys (Table 1).

4.2 Oscillation-Based Content Addressing Separates Similar Keys

ODN-SO(2) stores association i at read time t under $x_i = \hat{k}_i \mathcal{R}(H_t - H_i)^\top$, where $\mathcal{R}(a) = \bigoplus_r \exp(\omega_r a J)$ and $H_t = \sum_{j \le t} \Delta_j$ is the internal time (Appendix D). The overlap of two addresses then depends on the internal time between the writes:

$$G_{ij} = \hat{k}_i \mathcal{R}(H_i - H_j) \hat{k}_j^\top, \qquad C_u(a) = u \mathcal{R}(a) u^\top = \sum_{r=1}^{m} \eta_r \cos(\omega_r a), \tag{8}$$

where $C_u(a)$ is the overlap of a unit vector u with itself after internal time a, and $\eta_r = \|u_r\|_2^2$ is its energy in oscillator r. Without clock advance, $H_i = H_j$ and ODN reduces to Gated DeltaNet.

To quantify the separation, model the occurrences of one token as independent draws $\hat{k}_i = \bar{k} + \varepsilon_i$ around a mean $\bar{k} = \sqrt{c}\, u$, independent of the write times, so that c is their expected content overlap. For deviations ε_i that are isotropic on the complement of u, Proposition 2 gives

$$\mathbb{E}\, G_{ij}^2 \le c^2 C_u(H_i - H_j)^2 + \frac{1 - c^2}{d - 1}, \qquad \mathbb{E}\,(\hat{k}_i \hat{k}_j^\top)^2 \ge c^2, \tag{9}$$

and a general deviation covariance changes only the second term (Equation 27). Static content addressing keeps the shared overlap c^2. Oscillation-based content addressing multiplies it by C_u^2 for any similarity c, leaving a floor of order $1/d$; for unrelated keys, $c = 0$, the two address families coincide. The separation can be complete, with d copies of one key on orthonormal phase addresses (Proposition 3), but never exceeds the d dimensions of the state (Proposition 4). Section 5.1 tests this decomposition on a pretrained model.

4.3 The Spectrum Decides When to Correct and When to Retain

The kernel C_u decides whether a repeated key corrects its earlier association or coexists with it. Its shape is set by the spectrum and by the key's oscillator energies, and the clock maps token gaps to internal time. ODN keeps the log-uniform spectrum of RoPE (Equation 7). For frequencies spread log-uniformly over $[\omega_{\min}, \omega_{\max}]$, with $\Omega = \log(\omega_{\max}/\omega_{\min})$, the mean kernel has three regimes (Proposition 5). It stays near one for $a \lesssim \sqrt{\Omega}/\omega_{\max}$; it decreases linearly in $\log a$, by $\log 2/\Omega$ per doubling, for $1/\omega_{\max} \ll a \ll 1/\omega_{\min}$; and it vanishes beyond $1/\omega_{\min}$. This is the long-term decay of RoPE (Su et al., 2021). In a delta-rule memory it becomes an age-dependent update rule: content repeated within the short window is corrected, and content repeated across longer gaps is retained.

Table 1: Pretrained address geometry, averaged over 192 heads and 128 articles. (Appendix H.2).

	Gated DeltaNet	ODN-SO(2)	
Metric	Content address	Content address	Phase address
Mean cosine $\cos \downarrow$	0.270	0.287	0.106
Mean squared overlap $\cos^2 \downarrow$	0.155	0.167	0.077
Same-token squared overlap $\cos^2 \downarrow$	0.397	0.411	0.147
Same-token shared component c^2 or $c^2 C_u^2 \downarrow$	0.367	0.381	0.126
Participation ratio $(\sum_n \lambda_n)^2 / \sum_n \lambda_n^2 \uparrow$	11.1	10.8	28.4

5 EXPERIMENTS

We evaluate ODN at three levels. Pretrained SO(2) addresses and gate correlations test the mechanism; MQAR, Lag-Indexed Overwrite Recall, state tracking, and language modeling evaluate the complete models; and ablations isolate the shared clock and orbit geometry.

Setup. We compare ODN-SO(2) and ODN-SL(2) against Gated DeltaNet, KDA, and Selective RoPE implemented on a Gated DeltaNet backbone. All language models use matched 340M backbones and recurrent state sizes, and are trained on 15B FineWeb-Edu tokens (Penedo et al., 2024) with 4K contexts and identical optimizer settings. Synthetic tasks use smaller matched models: two layers of width 128 for MQAR and Lag-Indexed Overwrite Recall, and width 384 with one layer for S2 and two layers for A3. Full settings are in Appendix H.

5.1 INTERFERENCE AND ADDRESS SEPARATION

A delta write disturbs an earlier association in proportion to the overlap of their addresses, G_{ij} (Section 4.1). We measure these overlaps on 128 WikiText-2 articles of 4,096 tokens, comparing the content addresses of Gated DeltaNet and ODN-SO(2) with the phase addresses of ODN-SO(2) (Table 1); Appendix H.2 gives the details.

Gated DeltaNet keys are highly correlated. Gated DeltaNet keys are strongly correlated, both across tokens and between occurrences of the same token. Over all key pairs, the mean cosine is 0.270 and the mean squared overlap is 0.155, about twenty times the value $1/d$ expected for random keys, so the keys of a head share a common direction. For same-token pairs, the squared overlap rises to 0.397, and the shared component c^2 accounts for over 90% of it. Under static content addressing, this shared component makes delta writes interfere (Section 4.1).

Oscillation-based content addressing decorrelates keys and enlarges the address space. The content keys of ODN-SO(2) are as correlated as those of Gated DeltaNet (Table 1), so any decorrelation must come from the transport. As Equation 9 predicts, oscillation-based content addressing removes the shared component. Over all key pairs, the mean cosine falls from 0.287 to 0.106 and the mean squared overlap from 0.167 to 0.077. For same-token pairs, the squared overlap falls from 0.411 to 0.147. Nearly all of this drop is in the shared term, which falls from $c^2 = 0.381$ to $c^2 C_u^2 = 0.126$. The reduction grows with the gap between occurrences, following the kernel of Section 4.3 (Figure 2c). With the shared component removed, the addresses spread across the state: in the median head, the leading principal direction carries 11% of the variance instead of 28%, and the participation ratio, the effective number of address directions, rises from 10.8 to 28.4 (Figure 2a, b).

Table 2: ODN internal clock advance is correlated with write gate activation across evaluation corpora.

Task	Clock–write r
WikiText-2	$+0.322$
LAMBADA	$+0.301$
PIQA	$+0.335$
HellaSwag	$+0.282$
WinoGrande	$+0.285$
ARC-Easy	$+0.326$
ARC-Challenge	$+0.328$
BoolQ	$+0.360$

Clock–write coordination. The kernel is indexed by internal time rather than by token count: writes $j < i$ are separated by $H_i - H_j = \sum_{j<t\leq i} \Delta_t$, so where the clock advances decides which writes are pulled apart. We measure the

token-wise correlation between the clock step Δ_t and the write gate β_t on the eight evaluation corpora underlying Table 4, using 128 examples per task. The two are positively coupled on every task, with r between 0.28 and 0.36 (Table 2), so the learned clock tends to advance at the tokens that write, and internal time accumulates mainly across writes. Layer–head maps for WikiText-2, including the other gate combinations, appear in Appendix F.1.

5.2 Associative Recall

MQAR (Arora et al., 2024) tests associative recall with jointly learned keys, queries, and write controls. All methods succeed at low load and separate as load grows (Figure 3a). At the highest load, ODN-SL(2) leads KDA, ODN-SO(2) outperforms Gated DeltaNet and Selective RoPE, showing that oscillation-based content addressing improves learned recall over scalar transport and that learned orbit geometry provides a further gain at the same state size.

Lag-Indexed Overwrite Recall. Associative-recall benchmarks retrieve values through content keys (Danihelka et al., 2016; Arora et al., 2024), while recent repeated-key probes query only the latest of two bindings (Ghriss & Chakraborty, 2026). We instead test overwritten history across multiple writes under one key. Each example writes unique random values under that key, then provides a lag token $D \in [4, 16]$ and requests the value written D events earlier. For example,

Input	A	v0	A	v1	A	v2	A	v3	D=2
Output	∅	∅	∅	∅	∅	∅	∅	∅	v2

Table 3 shows that both ODN variants learn a delay-addressable overwrite history and transfer it to streams longer than those seen during training, while Gated DeltaNet, KDA, and Selective RoPE fail to recover the overwritten values. Removing clock advance collapses ODN-SO(2), identifying transport as the mechanism that creates the distinct phase addresses. Per-lag accuracies are in Appendix F.2.

5.3 State Tracking

State tracking tests whether a model can repeatedly apply a group operation while retaining the resulting state. We study two cyclic group tasks: S2 is running XOR (parity), and A3 is running addition modulo three (Grazzi et al., 2025; Movahedi et al., 2026).

Shared-clock rotations give ODN-SO(2) and ODN-SL(2) an inductive bias aligned with cyclic group composition. KDA uses positive diagonal retention, while Selective RoPE learns an independent phase drive at each frequency. On S2, both ODN variants reach 0.99 prefix-exact accuracy within 300 steps across all seeds, ahead of Selective RoPE and KDA (Figure 3b). On A3, ODN-SO(2) reaches sustained 0.99 mean accuracy first (Figure 3c).

5.4 Language Modeling

We evaluate the 340M models pretrained on FineWeb-Edu using perplexity on WikiText-2 and LAMBADA, and zero-shot accuracy on LAMBADA, PIQA, HellaSwag, WinoGrande, ARC-Easy, ARC-Challenge, and BoolQ (Table 4). ODN-SL(2) achieves the highest mean accuracy across the seven tasks, followed by ODN-SO(2).

ODN's clearest advantage is on LAMBADA: ODN-SL(2) achieves the lowest perplexity, and ODN-SO(2) achieves the highest accuracy. Predicting LAMBADA's final word draws on earlier context, consistent with the improved recall from oscillation-based content addressing. ODN-SO(2) also leads on HellaSwag and ARC-Easy, while ODN-SL(2) leads on WinoGrande and BoolQ. All four tasks select a response conditioned on the prompt: a plausible continuation, a referent, or an answer.

5.5 Transport Ablations

Table 5 isolates the contributions of the learned clock and spectrum, shared controls, and structured SL(2) generators.

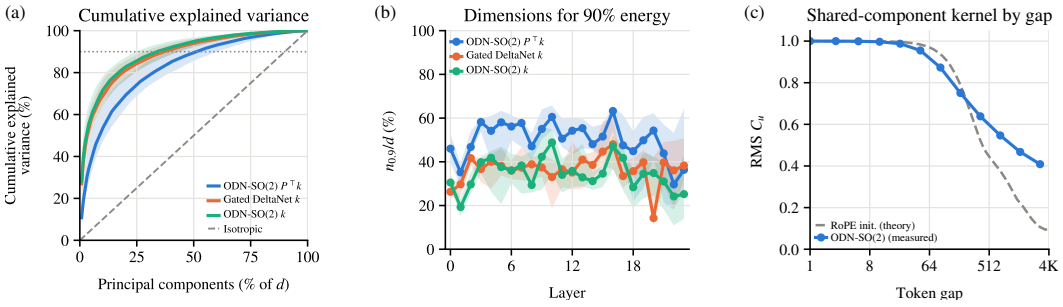

Figure 2: Write-address geometry on 128 WikiText-2 articles of 4,096 tokens. (a) Cumulative explained variance $E(n)$ of the top n principal components versus n/d, where $d = 128$. (b) Fraction $n_{0.9}/d$ of directions needed to capture 90% of the variance across layers. Lines show head-wise medians; bands show interquartile ranges. (c) Root mean square of C_u over same-token pairs of ODN-SO(2), by token gap, against the kernel of the RoPE-style initialization spectrum evaluated at each layer's mean learned clock rate and averaged over layers.

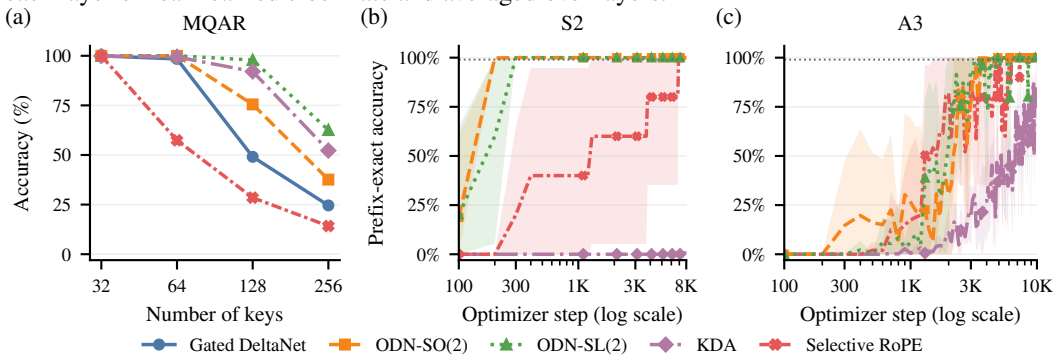

Figure 3: Synthetic benchmarks. (a) MQAR accuracy versus the number of key–value associations. (b, c) Prefix-exact state-tracking accuracy at length 128 on S2 and A3 over five seeds; bold curves and bands show the mean and one standard deviation, and the dotted line marks 0.99.

Table 3: Lag-Indexed Overwrite Recall accuracy (%) on 24–64 writes (in-distribution) and 80–128 writes (long horizon), 3,000 examples each. Best in bold, second best underlined.

Model	In-distribution	Long horizon
Gated DeltaNet	0.03	0.10
KDA	0.00	0.00
Selective RoPE	0.03	0.07
ODN-SO(2)	<u>99.17</u>	<u>99.23</u>
ODN-SL(2)	**100.00**	**99.93**
ODN-SO(2), no clock advance	3.43	4.07

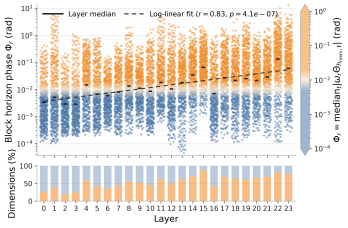

Figure 4: Oscillatory transport increases with layer depth. Top shows for each oscillator in the pretrained ODN-SO(2) the total rotation over the typical time period of each head over which memories persist. Bottom: Fraction of static (blue) vs rotating oscillators (orange) across layers. Deeper layers show systematically more rotation.

Adaptive clock and spectrum. Replacing the adaptive clock with a unit clock or freezing the spectrum worsens average accuracy, with the strongest decline when both changes are combined. Adaptive retention remains active throughout, so these comparisons show the contribution of adaptive oscillation-based content addressing. Replacing the log-uniform initial spectrum with odd harmonics of a base frequency also lowers average accuracy and raises LAMBADA perplexity, even though the spectrum remains learnable.

Shared versus independent controls. Giving each oscillator its own clock gate lowers average accuracy in both transport families; independent retention also offers no average-accuracy gain.

Table 4: Language modeling and zero-shot results at 340M parameters and 15B tokens. Avg. is the mean of the seven accuracy columns. Best in bold, second best underlined.

Model	Wiki. ppl↓	LMB. ppl↓	LMB. acc↑	PIQA acc↑	Hella. acc_n↑	Wino. acc↑	ARC-e acc↑	ARC-c acc_n↑	BoolQ acc↑	Avg. ↑
Gated DeltaNet	28.15	35.82	31.52	66.76	39.03	<u>51.54</u>	58.25	27.47	58.99	47.65
KDA	28.40	34.28	33.65	**67.14**	<u>39.66</u>	49.96	57.62	<u>27.90</u>	60.03	47.99
Selective RoPE	**27.86**	36.12	32.64	66.10	39.06	<u>51.54</u>	57.66	**28.67**	56.91	47.51
ODN-SO(2)	28.40	<u>32.46</u>	**34.02**	65.34	**39.72**	51.46	**59.22**	27.22	<u>60.70</u>	<u>48.24</u>
ODN-SL(2)	<u>28.02</u>	**32.23**	<u>33.67</u>	<u>66.97</u>	39.53	**52.64**	<u>58.63</u>	27.73	**60.89**	**48.58**

Table 5: Transport ablations at 340M parameters and 15B tokens. Independent gates act per oscillator; full scores are in Appendix F.3. Best in bold, second best underlined.

Parameterization	Wiki. ppl ↓	LMB. ppl ↓	Avg. acc ↑
ODN-SO(2)	28.40	<u>32.46</u>	<u>48.24</u>
Unit clock ($\Delta_t = 1$)	28.52	34.27	47.46
Fixed spectrum (ω_r frozen)	28.23	34.81	47.93
Unit clock + fixed spectrum	28.97	36.20	46.25
Odd-harmonic spectrum init	28.13	36.37	47.87
Independent clock gates	<u>27.69</u>	35.30	48.14
Independent retention	28.31	34.50	47.81
ODN-SL(2)	28.02	**32.23**	**48.58**
Independent clock gates	**27.62**	33.19	48.18
Independent retention	28.36	35.41	47.93
Generic ODN-SL(2)	29.24	34.60	48.02

Independent clock gates lower WikiText-2 perplexity but worsen LAMBADA perplexity, favoring a shared clock for recall-sensitive evaluation.

Generator parameterization. ODN-SL(2) lets each oscillator learn its own persistent orbit geometry under a shared clock. It achieves higher average accuracy than ODN-SO(2) with either independent clock gates or independent retention. The generic SL(2) parameterization, which learns rotation, squeeze, and shear rates independently, underperforms the oscillator-based SL(2). These comparisons favor learning structured orbit geometry over adding per-oscillator token-dependent controls. Appendix C.3 analyzes the orbit geometry learned by the pretrained ODN-SL(2).

5.6 Rotational Participation Increases with Layer Depth

To investigate if and how oscillatory transport is used across the pretrained ODN-SO(2) we quantified the total rotation of each oscillator over the typical internal time period U_{mem} over which associations decay by e^{-1} due to the retention gate. More concretely, for each oscillator r, we quantify the total rotation

$$\Theta_{\mathrm{mem},r} = |\omega_r U_{\mathrm{mem}}|$$

When $\Theta_{\mathrm{mem},r} \ll 1$ the block changes negligibly over the lifetime of the memory and is effectively static. We classify blocks with $\Theta_{\mathrm{mem}} \ll 10^{-2}$ as static and rotational otherwise. The figure shows that rotational participation increases with depth: deeper layers exhibit both larger total rotations and a greater fraction of rotational dimensions, consistent with an increasing role for phase-dependent dynamics at higher levels of representation.

6 Conclusion

Oscillatory Delta Networks rotate the recurrent state by a shared, content-dependent clock before each delta-rule write, so each association is addressed by both its content and its internal write time. This oscillation-based content addressing separates similar keys that static content addressing forces to overwrite: in a pretrained model it cuts same-token address overlap by 64%, and it lets ODN recover overwritten values where Gated DeltaNet, KDA, and Selective RoPE fail. A shared clock outperforms frequency-wise phase drives, and learning the orbit geometry adds further gains in recall and zero-shot accuracy. Structuring how a fixed-size memory transports what it holds is thus an effective alternative to enlarging it.

AI USE STATEMENT

We used generative AI tools (large language models) for three purposes. First, to aid and polish writing: editing the manuscript for clarity, grammar, and concision, and checking consistency among the text, tables, and notation. Second, to draft sections of the paper: the tools produced initial drafts of some passages, which the authors then revised and verified. Third, for retrieval and discovery: searching for and identifying related work, which the authors checked against the original sources. We did not use generative AI tools to formulate hypotheses, develop the theoretical results or proofs, design experiments, implement methods, generate or process data, or interpret results. We have reviewed all AI-assisted content, and we take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Simran Arora, Sabri Eyuboglu, Aman Timalsina, Isys Johnson, Michael Poli, James Zou, Atri Rudra, and Christopher Ré. Zoology: Measuring and Improving Recall in Efficient Language Models. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2312.04927.

Federico Barbero, Alex Vitvitskyi, Christos Perivolaropoulos, Razvan Pascanu, and Petar Veličković. Round and Round We Go! What Makes Rotary Positional Encodings Useful? In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.06205.

Jared Boyer, T. Konstantin Rusch, and Daniela Rus. Learning to Dissipate Energy in Oscillatory State-Space Models. *arXiv preprint arXiv:2505.12171*, 2025. URL https://arxiv.org/abs/2505.12171.

Rumen Dangovski, Li Jing, Preslav Nand, and Marin Soljačić. Rotational Unit of Memory: A Novel Representation Unit for RNNs with Scalable Applications. *Transactions of the Association for Computational Linguistics*, 7:121–138, 2019. URL https://arxiv.org/abs/1710.09537.

Ivo Danihelka, Greg Wayne, Benigno Uria, Nal Kalchbrenner, and Alex Graves. Associative Long Short-Term Memory. In *International Conference on Machine Learning*, 2016. URL https://arxiv.org/abs/1602.03032.

Tri Dao and Albert Gu. Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality. In *International Conference on Machine Learning*, 2024. URL https://arxiv.org/abs/2405.21060.

Ayoub Ghriss and Sourav Chakraborty. Kernelized Linear Attention: Breaking the Capacity Wall with Symmetric Cones. *arXiv preprint arXiv:2607.17419*, 2026. URL https://arxiv.org/abs/2607.17419.

Riccardo Grazzi, Julien Siems, Arber Zela, Jörg K. H. Franke, Frank Hutter, and Massimiliano Pontil. Unlocking State-Tracking in Linear RNNs Through Negative Eigenvalues. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2411.12537.

Albert Gu and Tri Dao. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. In *Conference on Language Modeling*, 2024. URL https://arxiv.org/abs/2312.00752.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently Modeling Long Sequences with Structured State Spaces. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2111.00396.

Andrew C. Heusser, David Poeppel, Youssef Ezzyat, and Lila Davachi. Episodic sequence memory is supported by a theta-gamma phase code. *Nature Neuroscience*, 19(10):1374–1380, 2016.

Marc W. Howard and Michael J. Kahana. A distributed representation of temporal context. *Journal of Mathematical Psychology*, 46(3):269–299, 2002.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention. In *International Conference on Machine Learning*, 2020. URL https://arxiv.org/abs/2006.16236.

Kimi Team. Kimi Linear: An Expressive, Efficient Attention Architecture. *arXiv preprint arXiv:2510.26692*, 2025. URL https://arxiv.org/abs/2510.26692.

Teuvo Kohonen. Correlation Matrix Memories. *IEEE Transactions on Computers*, C-21(4):353–359, 1972. doi: 10.1109/TC.1972.5008975.

Aakash Lahoti, Kevin Y. Li, Berlin Chen, Caitlin Wang, Aviv Bick, J. Zico Kolter, Tri Dao, and Albert Gu. Mamba-3: Improved Sequence Modeling using State Space Principles. *arXiv preprint arXiv:2603.15569*, 2026. URL https://arxiv.org/abs/2603.15569.

John E. Lisman and Marco A. P. Idiart. Storage of 7 $\pm$ 2 short-term memories in oscillatory subcycles. *Science*, 267(5203):1512–1515, 1995.

John E. Lisman and Ole Jensen. The theta-gamma neural code. *Neuron*, 77(6):1002–1016, 2013.

Emily A. Mankin, Fraser T. Sparks, Begum Slayyeh, Robert J. Sutherland, Stefan Leutgeb, and Jill K. Leutgeb. Neuronal code for extended time in the hippocampus. *Proceedings of the National Academy of Sciences*, 109(47):19462–19467, 2012.

Sajad Movahedi, Timur Carstensen, Arshia Afzal, Frank Hutter, Antonio Orvieto, and Volkan Cevher. Selective Rotary Position Embedding. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2511.17388.

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*, 2023. URL https://arxiv.org/abs/2303.06349.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro von Werra, and Thomas Wolf. The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2406.17557.

Tony A. Plate. Holographic Reduced Representations. *IEEE Transactions on Neural Networks*, 6 (3):623–641, 1995.

Alon Rubin, Nitzan Geva, Liron Sheintuch, and Yaniv Ziv. Hippocampal ensemble dynamics timestamp events in long-term memory. *eLife*, 4:e12247, 2015.

T. Konstantin Rusch and Daniela Rus. Oscillatory State-Space Models. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.03943.

Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. Linear Transformers Are Secretly Fast Weight Programmers. In *International Conference on Machine Learning*, 2021. URL https://arxiv.org/abs/2102.11174.

Jürgen Schmidhuber. Learning to Control Fast-Weight Memories: An Alternative to Dynamic Recurrent Networks. *Neural Computation*, 4(1):131–139, 1992. URL https://people.idsia.ch/~juergen/fastweights/.

Markus Siegel, Melissa R. Warden, and Earl K. Miller. Phase-dependent neuronal coding of objects in short-term memory. *Proceedings of the National Academy of Sciences*, 106(50):21341–21346, 2009.

Julien Siems, Timur Carstensen, Arber Zela, Frank Hutter, Massimiliano Pontil, and Riccardo Grazzi. DeltaProduct: Improving State-Tracking in Linear RNNs via Householder Products. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2502.10297.

Prajwal Singhania, Siddharth Singh, Shwai He, Soheil Feizi, and Abhinav Bhatele. Loki: Low-rank keys for efficient sparse attention. *Advances in Neural Information Processing Systems*, 37: 16692–16723, 2024.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced Transformer with Rotary Position Embedding. *arXiv preprint arXiv:2104.09864*, 2021. URL https://arxiv.org/abs/2104.09864.

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. URL https://arxiv.org/abs/2307.08621.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, 2017. URL https://arxiv.org/abs/1706.03762.

Songlin Yang, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. Parallelizing Linear Transformers with the Delta Rule over Sequence Length. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2406.06484.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated Delta Networks: Improving Mamba2 with Delta Rule. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2412.06464.

A NOTATION

Table 6 lists the symbols that recur across sections. Keys and queries are row vectors throughout, and the memory is read as kS. Within a head, r indexes oscillators, i and t index tokens or stored associations, and N counts stored associations.

Table 6: Symbols used in more than one section.

Symbol	Meaning	Defined in
Sizes		
$d = 2m$	address and state dimension	Section 3.1
m	oscillators per head	Section 3.1
d_v	value dimension	Section 3.1
L	sequence length	Section 3.2
b	chunk width ($b = 16$)	Appendix G
Layer		
k_t, q_t, v_t	key, query, value (row vectors)	Section 3.1
S_t	memory	Equation 2
$\bar{S}_t$	transported memory	Equation 2
e_t	prediction residual	Equation 15
β_t	write gate	Section 3.1
ρ_t	retention per head	Section 3.1
$\kappa_{t,r}$	hyperbolic guard multiplier	Equation 6
Δ_t	shared clock step	Equation 3
$H_t = \sum_{j \le t} \Delta_j$	internal time	Section 4
Transport		
ω_r	learned frequency of oscillator r	Section 3.2
A_r	persistent generator of oscillator r	Equation 5
$T_t, \Phi_{t \leftarrow i}$	one-step transition, interval transport	Equation 14
$\mathcal{R}(a)$	block rotation by internal age a	Section 4
ϕ_r, ψ_r	SL(2) shear and squeeze parameters	Equation 5
$\delta_r = 1 - \phi_r^2 - \psi_r^2$	discriminant; its sign selects the branch	Equation 5
ω_r^+	hyperbolic guard rate $\omega_r \sqrt{\max(-\delta_r, 0)}$	Equation 6
Addresses		
x_i	unit address in the read frame	Equation 19
$G = XX^\top$	address Gram matrix	Equation 8
$C_u(a)$	self-overlap of u after internal time a	Equation 8
η_r	oscillator energy of a unit vector	Equation 8
$\bar{k}, c, \Sigma$	mean key, $\|\bar{k}\|_2^2$, deviation covariance	Equation 26
Ω	log bandwidth $\log(\omega_{\max}/\omega_{\min})$ of the spectrum	Section 4.3
$\bar{C}(a)$	mean kernel of a log-uniform spectrum	Equation 31
Chunkwise algorithm		
$\mathcal{B} = \mathrm{Diag}(\beta_i)$	chunk write gates	Equation 42
Ψ_{kk}, Ψ_{qk}	key–key and query–key interactions	Equation 41
W, U, M	WY auxiliaries and their triangular factor	Equation 43

B EXTENDED RELATED WORK

S4 introduced efficient structured state-space parameterizations for long-range dependencies (Gu et al., 2022), and the Linear Recurrent Unit showed that diagonal complex recurrences with suitable initialization and normalization are effective on long sequences (Orvieto et al., 2023). ODN instead places oscillatory transport inside a delta-rule fast-weight memory. The Rotational Unit of Memory uses learned rotations as its state transition to preserve norms over long horizons (Dangovski et al., 2019), and Associative LSTM stores redundant copies of a complex-valued memory under random phase rotations to reduce retrieval noise (Danihelka et al., 2016); in ODN the phase comes from a learned clock, so a single memory copy is addressed by write time.

Barbero et al. (2025) show that trained RoPE models concentrate query and key norm on particular frequencies, using high frequencies for positional attention patterns and low frequencies for content matching that is insensitive to distance. Appendix D gives the exact correspondence between unit-clock Hebbian ODN and rotary linear attention. Our S2 and A3 experiments examine how quickly the tested parameterizations learn simple group computations; they do not establish expressivity for arbitrary state-tracking problems (Grazzi et al., 2025; Siems et al., 2025).

C Explicit Transport Matrices and SL(2) Flows

This appendix gives the full SO(2) transport matrix and the exact SL(2) flow used to implement Section 3.

C.1 ODN-SO(2): Rotational Transport

In oscillator ordering, the full transport for one head is

$$T_t = T_t^{\mathrm{SO}(2)} = \rho_t \begin{bmatrix} \cos\theta_{t,1} & -\sin\theta_{t,1} & & & \\ \sin\theta_{t,1} & \cos\theta_{t,1} & & & \\ & & \ddots & & \\ & & & \cos\theta_{t,m} & -\sin\theta_{t,m} \\ & & & \sin\theta_{t,m} & \cos\theta_{t,m} \end{bmatrix}, \tag{10}$$

$$T_t = \rho_t \bigoplus_{r=1}^{m} \exp(\omega_r \Delta_t J), \qquad \bar{S}_t = T_t^{\mathrm{SO}(2)} S_{t-1},$$

where $\theta_{t,r} = \omega_r \Delta_t$ is the phase advance of oscillator r at token t. The implementation applies the 2×2 blocks directly without materializing the $d \times d$ matrix.

C.2 ODN-SL(2): Learned Flow Geometry

The three matrices multiplying ω_r, $\omega_r \phi_r$, and $\omega_r \psi_r$ in Equation 5 span $\mathfrak{sl}(2, \mathbb{R})$. Since A_r is traceless, Cayley–Hamilton gives $A_r^2 = -(\det A_r) I$ and

$$\exp(\Delta_t A_r) = \mathcal{C}(\xi_{t,r})\, I + \Delta_t\, \mathcal{S}(\xi_{t,r})\, A_r, \qquad \xi_{t,r} = \Delta_t^2 \det A_r, \tag{11}$$

$$\mathcal{C}(\xi) = \sum_{n \geq 0} \frac{(-\xi)^n}{(2n)!}, \qquad \mathcal{S}(\xi) = \sum_{n \geq 0} \frac{(-\xi)^n}{(2n+1)!}. \tag{12}$$

Both series are entire, so Equation 11 is smooth in δ_r across the sign change; for $\xi_{t,r} > 0$ they evaluate to $\cos\sqrt{\xi_{t,r}}$ and $\sin\sqrt{\xi_{t,r}}/\sqrt{\xi_{t,r}}$, and for $\xi_{t,r} < 0$ to $\cosh\sqrt{-\xi_{t,r}}$ and $\sinh\sqrt{-\xi_{t,r}}/\sqrt{-\xi_{t,r}}$. For $\delta_r > 0$, the flow is a rotation at frequency $\omega_r\sqrt{\delta_r}$ in a skewed frame; $\delta_r = 1$ recovers SO(2). For $\delta_r = 0$, A_r is nilpotent and $\exp(\Delta_t A_r) = I + \Delta_t A_r$ is a parabolic shear. For $\delta_r < 0$, the eigenvalues are $\pm\omega_r\sqrt{-\delta_r}$ and the flow is a hyperbolic squeeze. Figure 5 illustrates the four regimes.

The flow is unimodular and obeys $\exp(\Delta A_r)\exp(\Delta' A_r) = \exp((\Delta + \Delta')A_r)$, which permits separate associative scans of the determinant-one flow and its decay in Appendix G. For $\delta_r < 0$, the guard uses the spectral abscissa ω_r^+ because $\sum_t \omega_r^+ \Delta_t = \omega_r^+ \sum_t \Delta_t$. In the eigenframe Π_r of A_r, the guarded transition is

$$T_{t,r} = \rho_t \Pi_r \,\mathrm{diag}\Big(1, e^{-2\omega_r^+ \Delta_t}\Big)\, \Pi_r^{-1}. \tag{13}$$

Thus the expanding eigenmode retains only the base retention, while the other contracts at twice the hyperbolic rate. For $\delta_r \geq 0$, $\omega_r^+ = 0$ and $\kappa_{t,r} = 1$.

C.3 Learned SL(2) Geometry in the Pretrained Model

ODN-SL(2) uses its additional geometric freedom primarily to reshape oscillatory orbits. Of the 12,288 oscillators in the pretrained 340M model, 98.45% remain in the elliptic regime with $\delta_r > 10^{-3}$ (Figure 6). The deformation grows across depth groups: mean δ_r falls from 0.951 in layers 0–7 to 0.860 in layers 8–15 and 0.838 in layers 16–23, compared with $\delta_r = 1$ at the circular

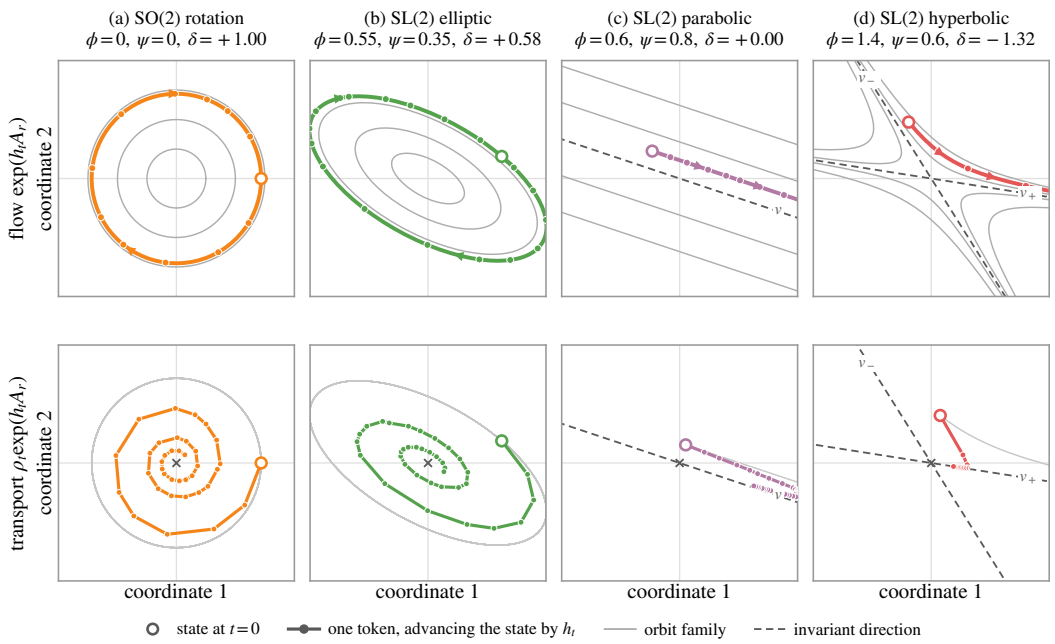

Figure 5: One-oscillator trajectories at $\omega_r = 0.9$. The top row shows the determinant-one flow; the bottom row adds retention with $\rho_t = e^{-0.055}$. All panels use the same clock sequence. Columns show (a) circular SO(2), (b) skewed elliptic, (c) parabolic, and (d) guarded hyperbolic dynamics. Axis ranges vary by column. The learned (ϕ_r, ψ_r) distribution appears in Figure 6.

initialization. Training therefore preserves predominantly oscillatory dynamics while adapting the orbits more strongly in deeper layers.

For an elliptic oscillator, the learned flow is a rotation at frequency $\omega_r\sqrt{\delta_r}$ in a learned skewed coordinate frame (Appendix C.2, above). This frame turns circular trajectories into ellipses and changes the relative weighting of directions in address overlap. SL(2) can thus adapt both the shape of a phase trajectory and its effective frequency, while the shared clock continues to advance every oscillator along one internal-time coordinate. The small hyperbolic subset additionally acquires directional retention through the guard in Equation 6.

The geometry analysis and transport ablations support a common design principle: learn the spatial geometry of oscillation-based content addressing while retaining one shared temporal coordinate per head. ODN-SL(2) realizes this principle as a predominantly oscillatory memory whose address geometry is learned alongside its clock and spectrum.

D Phase-Addressed Memory Expansion

This section expands the transported recurrence for both ODN variants and then specializes the address geometry to SO(2).

Using oscillator ordering, assemble T_t from the blocks of Equations 3 and 6. The transport from write time i to read time t is

$$\Phi_{t\leftarrow i} = T_t T_{t-1} \cdots T_{i+1}, \qquad \Phi_{t\leftarrow t} = I. \tag{14}$$

For $S_0 = 0$, Equation 2 gives

$$S_t = \sum_{i=1}^{t} \Phi_{t\leftarrow i}\, \beta_i k_i^\top e_i, \qquad e_i = v_i - k_i \bar{S}_i. \tag{15}$$

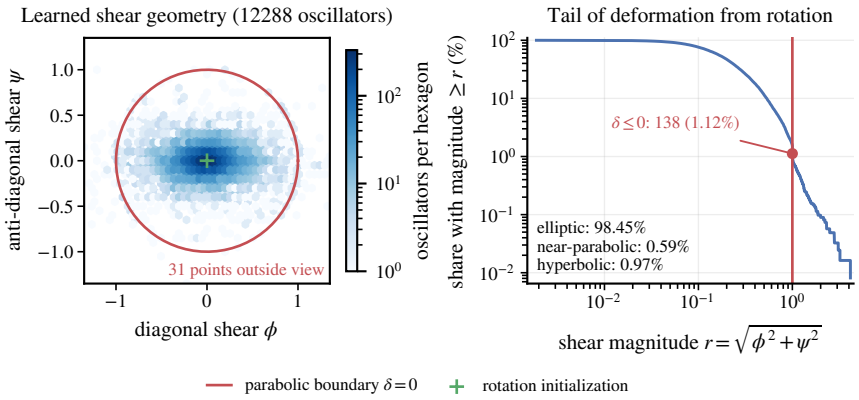

Figure 6: Learned SL(2) geometry in the 340M model. Left: density of oscillators in the (ϕ_r, ψ_r) plane on a logarithmic color scale; the origin is the rotation initialization and the unit circle separates elliptic from hyperbolic flows. Right: empirical tail of deformation magnitude $\varrho_r = \sqrt{\phi_r^2 + \psi_r^2}$. Branch percentages use the band $|\delta_r| \le 10^{-3}$.

Reading with the current query yields

$$o_t = q_t S_t = \sum_{i=1}^{t} \beta_i \left(q_t \Phi_{t\leftarrow i} k_i^\top\right) e_i. \tag{16}$$

Association i is therefore addressed by its transported key $k_i \Phi_{t\leftarrow i}^\top$. For Gated DeltaNet, $\Phi_{t\leftarrow i} = \gamma_{t\leftarrow i} I$ with $\gamma_{t\leftarrow i} = \prod_{j=i+1}^{t} \rho_j$, which retains the content address up to a scalar retention factor. For ODN-SO(2), $\Phi_{t\leftarrow i} = \gamma_{t\leftarrow i} \mathcal{R}(H_t - H_i)$, and the coefficient of residual e_i becomes

$$q_t \Phi_{t\leftarrow i} k_i^\top = \gamma_{t\leftarrow i} \sum_{r=1}^{m} \|q_{t,r}\| \, \|k_{i,r}\| \cos\!\left(\alpha_{t,r}^q - \alpha_{i,r}^k - \omega_r (H_t - H_i)\right), \tag{17}$$

where $\alpha_{t,r}^q$ and $\alpha_{i,r}^k$ are the polar angles of the query and key components. Each oscillator combines content orientation and internal age in one score. For two transported keys, the same expansion gives their overlap; a later common SO(2) rotation cancels, so it depends only on the retention factors and the internal time $H_i - H_j$ between the writes (Equation 8).

Relation to RoPE linear attention. With the unit clock $\Delta_t = 1$, unit retention, and Hebbian writes ($e_i = v_i$), $\Phi_{t\leftarrow i} = \mathcal{R}(t-i)$, and Equation 16 becomes

$$o_t = \sum_{i\le t} \beta_i \left(q_t \mathcal{R}(t-i) k_i^\top\right) v_i = \sum_{i\le t} \beta_i \left\langle q_t \mathcal{R}(t),\, k_i \mathcal{R}(i)\right\rangle v_i, \tag{18}$$

by the group identity $\mathcal{R}(t)\mathcal{R}(i)^\top = \mathcal{R}(t-i)$. With the row-vector convention $\mathrm{RoPE}_n(x) = x\mathcal{R}(n)$, this is exactly causal, unnormalized linear attention with a RoPE inner-product kernel (Su et al., 2021). Softmax attention uses the same rotary score inside a normalized exponential readout, which this finite-state recurrence does not reproduce. Token-dependent steps and scalar retention replace the score by $\gamma_{t\leftarrow i}\langle q_t \mathcal{R}(H_t), k_i \mathcal{R}(H_i)\rangle$, a rotary kernel on a non-uniform time grid shared by the head's frequencies. Delta-rule writing keeps this kernel and replaces each raw value v_i by the residual e_i.

E Proofs for Similar Keys and Oscillation-Based Content Addressing

This appendix proves the statements of Section 4.

The results concern orthogonal rotation. Two steps use orthogonality directly: transport keeps addresses on the unit sphere, and a common change of read time leaves the Gram matrix invariant.

ODN-SL(2) transport preserves neither exactly, so the corresponding statements for it would also have to control the metric distortion of its learned flow (Appendix C.2).

E.1 Setting and Error Identities

Consider N associations written at internal times $H_1 \le \cdots \le H_N$ and read at internal time $H_t \ge H_N$, indexed chronologically. Retention is one, each of the N writes has unit strength, and no other write occurs, so between two of them the state is only transported. Keys are unit vectors $\hat{k}_i \in \mathbb{R}^{1\times d}$, $d = 2m$. The content and phase addresses are

$$x_i^{\text{content}} = \hat{k}_i, \qquad x_i^{\text{phase}} = \hat{k}_i \mathcal{R}(H_t - H_i)^\top, \qquad \|x_i^{\text{content}}\|_2 = \|x_i^{\text{phase}}\|_2 = 1, \tag{19}$$

and association i is read with the matched query $q = x_i$, so $\widehat{v}_i = x_i S_N$. Let X stack the addresses as rows and $G = XX^\top$. The values are zero mean, satisfy $\mathbb{E}\langle v_i, v_j\rangle = \delta_{ij}$, and are independent of the addresses. Stacking them as the rows of V, any coefficient row vector ζ that depends only on the addresses satisfies

$$\mathbb{E}\|\zeta V\|_2^2 = \|\zeta\|_2^2. \tag{20}$$

The worst-item recall error is

$$\mathcal{E}_N = \max_{i\le N} \mathbb{E}\|\widehat{v}_i - v_i\|_2^2, \tag{21}$$

and the zero readout has error one.

Hebbian storage. With $S_N = X^\top V$, the matched read is $\widehat{v}_i = \sum_j G_{ij} v_j$, and Equation 20 gives

$$\mathbb{E}\|\widehat{v}_i - v_i\|_2^2 = \sum_{j\ne i} G_{ij}^2. \tag{22}$$

Delta-rule writing in the read frame. Let S_i be the state after write i and $\tilde{S}_i = \mathcal{R}(H_t - H_i)S_i$ the same state expressed in the frame of the final read. The transported memory before write i is $\bar{S}_i = \mathcal{R}(H_i - H_{i-1})S_{i-1} = \mathcal{R}(H_i - H_t)\tilde{S}_{i-1}$, so the residual is $e_i = v_i - \hat{k}_i\bar{S}_i = v_i - x_i\tilde{S}_{i-1}$, and $\tilde{S}_i = \tilde{S}_{i-1} + x_i^\top e_i$. Dropping the tilde, delta-rule writing is therefore

$$S_i = S_{i-1} + x_i^\top e_i, \qquad e_i = v_i - x_i S_{i-1}, \qquad S_0 = 0, \tag{23}$$

for either address family. Let E stack the residuals and let Λ be the strictly lower-triangular part of $G = I + \Lambda + \Lambda^\top$. Because $x_i S_{i-1} = \sum_{j<i} G_{ij} e_j$, the residuals satisfy $(I + \Lambda)E = V$; then $S_N = X^\top E$ and $XS_N = GE = V + \Lambda^\top E$, so

$$E = (I + \Lambda)^{-1}V, \qquad S_N = X^\top E, \qquad XS_N - V = \Lambda^\top (I + \Lambda)^{-1}V. \tag{24}$$

Row i of the last identity is the recall error of association i. The erasure itself is visible in Equation 23: the earlier association i contributes $x_i^\top e_i$ to S_{j-1}, and the term $-x_j^\top x_j S_{j-1}$ of write j removes $G_{ji}\, x_j^\top e_i$ from it. Write j therefore changes the stored contribution of association i in proportion to G_{ji}, the overlap that Section 4 studies.

E.2 Static Content Addressing with Similar Keys

Proposition 1 (similar keys overwrite each other)**.** *Suppose $|G_{ij}| \ge c_0 > 0$ for all distinct $i, j \le N$. Then for $N \ge 2$,*

$$\mathcal{E}_N^{\text{Hebb}} \ge (N-1)c_0^2, \qquad \mathcal{E}_N^{\text{delta}} \ge c_0^2. \tag{25}$$

If all keys are identical, then $\mathcal{E}_N^{\text{delta}} = 2$.

Proof. The Hebbian bound follows from Equation 22. For the delta rule, consider association $N-1$. Row $N-1$ of the strictly upper-triangular $\Lambda^\top$ has a single nonzero entry, $G_{N,N-1}$, in column N, so row $N-1$ of $\Lambda^\top(I+\Lambda)^{-1}$ equals $G_{N,N-1}$ times row N of $(I+\Lambda)^{-1}$. That matrix is unit lower-triangular, so its row N has norm at least one, and Equations 20 and 24 give an error of at least $G_{N,N-1}^2 \ge c_0^2$ on association $N-1$. For identical keys, $G = \mathbf{1}\mathbf{1}^\top$, the residuals are $e_1 = v_1$ and $e_i = v_i - v_{i-1}$, so $S_N = \hat{k}^\top v_N$ and every earlier association reads v_N, with error $\mathbb{E}\|v_N - v_i\|_2^2 = 2$. □

For content addresses, $G_{ij} = \hat{k}_i \hat{k}_j^\top$ is fixed by the keys alone, so neither bound depends on d or on when the writes occurred. For two keys with overlap c_0, the older value has delta-rule error $c_0^2 + c_0^4$, larger than the Hebbian c_0^2: the delta rule makes the newest association exact at the expense of the older one.

E.3 Phase-Address Overlap and the Shared Component

Since $\mathcal{R}(a)^\top = \mathcal{R}(-a)$, the phase addresses satisfy $G_{ij}^{\text{phase}} = \hat{k}_i \mathcal{R}(H_i - H_j)\hat{k}_j^\top$, which does not depend on the read time and reduces to the content overlap when $H_i = H_j$. For a unit vector u with oscillator energies $\eta_r = \|u_r\|_2^2$, $\sum_r \eta_r = 1$, each 2×2 block contributes $u_r \exp(\omega_r a J) u_r^\top = \eta_r \cos(\omega_r a)$ because $u_r J u_r^\top = 0$, which gives the kernel $C_u(a)$ of Equation 8.

Proposition 2 (shared-component identity)**.** *Let the unit keys be independent draws with mean $\bar{k} = \mathbb{E}\hat{k}_i$ and covariance $\Sigma = \mathbb{E}(\hat{k}_i - \bar{k})^\top(\hat{k}_i - \bar{k})$, independent of the write times, and let $c = \|\bar{k}\|_2^2 > 0$ and $u = \bar{k}/\sqrt{c}$. For $i \neq j$, write $a = H_i - H_j$ and $R_a = \mathcal{R}(a)$. Then*

$$\begin{aligned} \mathbb{E}\,(G_{ij}^{\text{phase}})^2 &= c^2 C_u(a)^2 + \bar{k} R_a \Sigma R_a^\top \bar{k}^\top + \bar{k} R_a^\top \Sigma R_a \bar{k}^\top + \operatorname{tr}\big(R_a \Sigma R_a^\top \Sigma\big), \\ \mathbb{E}\,(G_{ij}^{\text{content}})^2 &= c^2 + 2\bar{k}\Sigma\bar{k}^\top + \operatorname{tr}\Sigma^2 \;\geq\; c^2, \end{aligned} \tag{26}$$

and consequently

$$\mathbb{E}\,(G_{ij}^{\text{phase}})^2 \leq c^2 C_u(a)^2 + 2c\,\lambda_{\max}(\Sigma) + \operatorname{tr}\Sigma^2. \tag{27}$$

If the deviation is isotropic on the complement of u, that is $\Sigma = (1-c)(I - u^\top u)/(d-1)$, then

$$\mathbb{E}\,(G_{ij}^{\text{phase}})^2 = c^2 C_u^2 + \frac{2c(1-c)(1-C_u^2)}{d-1} + \frac{(1-c)^2(d-2+C_u^2)}{(d-1)^2} \leq c^2 C_u(a)^2 + \frac{1-c^2}{d-1}. \tag{28}$$

Proof. Write $\hat{k}_i = \bar{k} + \varepsilon_i$ with independent zero-mean deviations. Then $G_{ij}^{\text{phase}} = \bar{k} R_a \bar{k}^\top + \bar{k} R_a \varepsilon_j^\top + \varepsilon_i R_a \bar{k}^\top + \varepsilon_i R_a \varepsilon_j^\top$, with $\bar{k} R_a \bar{k}^\top = c\, C_u(a)$. Every product of two distinct terms contains a deviation that appears once and is independent of the rest, so its expectation vanishes. The squared terms have expectations $\bar{k} R_a \Sigma R_a^\top \bar{k}^\top$, $\bar{k} R_a^\top \Sigma R_a \bar{k}^\top$, and $\mathbb{E}\operatorname{tr}(R_a \varepsilon_j^\top \varepsilon_j R_a^\top \varepsilon_i^\top \varepsilon_i) = \operatorname{tr}(R_a \Sigma R_a^\top \Sigma)$, which proves Equation 26; setting $R_a = I$ gives the content case. Since R_a is orthogonal, $\bar{k} R_a \Sigma R_a^\top \bar{k}^\top \leq c\lambda_{\max}(\Sigma)$, and Cauchy–Schwarz gives $\operatorname{tr}(R_a \Sigma R_a^\top \Sigma) \leq \operatorname{tr}\Sigma^2$. In the isotropic case, $\bar{k} R_a (I - u^\top u) R_a^\top \bar{k}^\top = c(1 - C_u^2)$ for both middle terms, and with $\tilde{u} = u R_a^\top$, $\operatorname{tr}\big((I - \tilde{u}^\top \tilde{u})(I - u^\top u)\big) = d - 2 + C_u^2$. The final inequality uses $1 - C_u^2 \leq 1$ and $d - 2 + C_u^2 \leq d - 1$. □

For unrelated keys, $c = 0$, the two address families coincide: for a fixed spectrum, independent isotropic unit keys, and write times independent of those keys, $\hat{k}_i \mathcal{R}(H_i - H_j)$ is again isotropic and independent of $\hat{k}_j$, so both have $\mathbb{E}|G_{ij}|^2 = 1/d$. Under Hebbian storage, Equation 22 turns these expectations into expected recall error; under delta-rule writing they are the expected squared erasure coefficients of Equation 24.

E.4 Exact Separation and the Dimension Limit

For a spacing $\tau > 0$, choose

$$\omega_r = \frac{(2r-1)\pi}{d\tau}, \quad r = 1, \ldots, m, \qquad H_i = H_1 + (i-1)\tau, \quad i = 1, \ldots, d. \tag{29}$$

Proposition 3 (one content key, d orthogonal phase addresses)**.** *For a unit key with uniform oscillator energies $\eta_r = 1/m$ and the spectrum and write times of Equation 29, the d phase addresses form an orthonormal basis of $\mathbb{R}^d$, while the content addresses coincide. Under both Hebbian and delta-rule writing, oscillation-based content addressing recalls all d values exactly, whereas static content addressing has $\mathcal{E}_N \geq 1$ for every $2 \leq N \leq d$.*

Proof. Let $u = \hat{k}$. For any integer ℓ with $0 < |\ell| < d$, the finite cosine-sum identity $\sum_{r=1}^{m} \cos((2r-1)\theta) = \sin(2m\theta)/(2\sin\theta)$ at $\theta = \pi\ell/d$ gives

$$C_u(\ell\tau) = \frac{1}{m}\sum_{r=1}^{m} \cos\left(\frac{(2r-1)\pi\ell}{d}\right) = \frac{\sin(\pi\ell)}{d\sin(\pi\ell/d)} = 0. \tag{30}$$

With $C_u(0) = 1$, $G^{\text{phase}} = I_d$ and $\Lambda = 0$, so Equations 22 and 24 give zero error for both write rules. The content addresses have $G^{\text{content}} = \mathbf{1}\mathbf{1}^\top$, and Proposition 1 with $c_0 = 1$ gives Hebbian error $N - 1$ and delta-rule error 2. □

The code is confined to a finite age window: $C_u(d\tau) = -1$, so a further write on the same grid collides with the first at full magnitude. Oscillation-based content addressing cannot exceed the state dimension.

Proposition 4 (dimension limit)**.** *For any addresses independent of the values and any write rule, a matched linear readout of a $d \times d_v$ state satisfies $\mathcal{E}_N \geq 1 - d/N$.*

Proof. The columns of XS_N lie in the column space of X, whose orthogonal projector $P_X \in \mathbb{R}^{N\times N}$ has rank at most d and depends only on the addresses. Whatever S_N is, including a state that depends on the values, $\|XS_N - V\|_F^2 \geq \|(I - P_X)V\|_F^2$. Since $\mathbb{E}VV^\top = I_N$, $\mathbb{E}\|(I - P_X)V\|_F^2 = \operatorname{tr}(I - P_X) \geq N - d$, and the worst item error is at least the average, $(N - d)/N$. □

E.5 The Kernel of a RoPE-Style Spectrum

Let $\operatorname{Ci}(x) = -\int_x^\infty t^{-1}\cos t\, dt$ denote the cosine integral and γ_{E} the Euler–Mascheroni constant.

Proposition 5 (RoPE-style kernel)**.** *Draw $\log\omega_r$ independently and uniformly on $[\log\omega_{\min}, \log\omega_{\max}]$, let $\Omega = \log(\omega_{\max}/\omega_{\min})$, and fix a unit u with energies η_r and $m_{\text{eff}} = 1/\sum_r \eta_r^2$. The kernel of Equation 8 has mean and variance*

$$\bar{C}(a) = \mathbb{E}_\omega C_u(a) = \frac{\operatorname{Ci}(\omega_{\max}a) - \operatorname{Ci}(\omega_{\min}a)}{\Omega}, \qquad \operatorname{Var}_\omega C_u(a) \leq \frac{1}{m_{\text{eff}}}, \tag{31}$$

and for every $a > 0$,

$$\bar{C}(a) \geq 1 - \frac{(\omega_{\max}^2 - \omega_{\min}^2)\, a^2}{4\Omega}, \tag{32}$$

$$\left|\bar{C}(a) - \frac{\log\frac{1}{\omega_{\min}a} - \gamma_{\text{E}}}{\Omega}\right| \leq \frac{1}{\Omega}\left(\frac{2}{\omega_{\max}a} + \frac{(\omega_{\min}a)^2}{4}\right), \tag{33}$$

$$|\bar{C}(a)| \leq \frac{4}{\Omega\,\omega_{\min}a}. \tag{34}$$

Proof. Each ω_r has density $(\Omega\omega)^{-1}$ on $[\omega_{\min}, \omega_{\max}]$, so $\mathbb{E}\cos(\omega_r a) = \Omega^{-1}\int_{\omega_{\min}a}^{\omega_{\max}a} t^{-1}\cos t\, dt$, which is Equation 31 for every r and hence for any weights η_r. The summands are independent and bounded by η_r in magnitude, so the variance is at most $\sum_r \eta_r^2$. For Equation 32, $\cos x \geq 1 - x^2/2$ and $\mathbb{E}\omega_r^2 = (\omega_{\max}^2 - \omega_{\min}^2)/(2\Omega)$. For the other two, write $\operatorname{Ci}(x) = \gamma_{\text{E}} + \log x + \int_0^x t^{-1}(\cos t - 1)\, dt$, whose last term is at most $x^2/4$ in magnitude because $|\cos t - 1| \leq t^2/2$, and integrate by parts to obtain $\operatorname{Ci}(x) = x^{-1}\sin x - \int_x^\infty t^{-2}\sin t\, dt$, so $|\operatorname{Ci}(x)| \leq 2/x$. Substituting the first form for $\operatorname{Ci}(\omega_{\min}a)$ and the bound for $\operatorname{Ci}(\omega_{\max}a)$ gives Equation 33; the bound for both gives Equation 34, using $\omega_{\min} \leq \omega_{\max}$. □

F Supplementary Experiments and Analysis

This appendix supplies the measurements and controls behind Section 5; it does not restate the results reported there. The models, budgets, and evaluation harness are those of Section 5, specified in full in Appendix H.

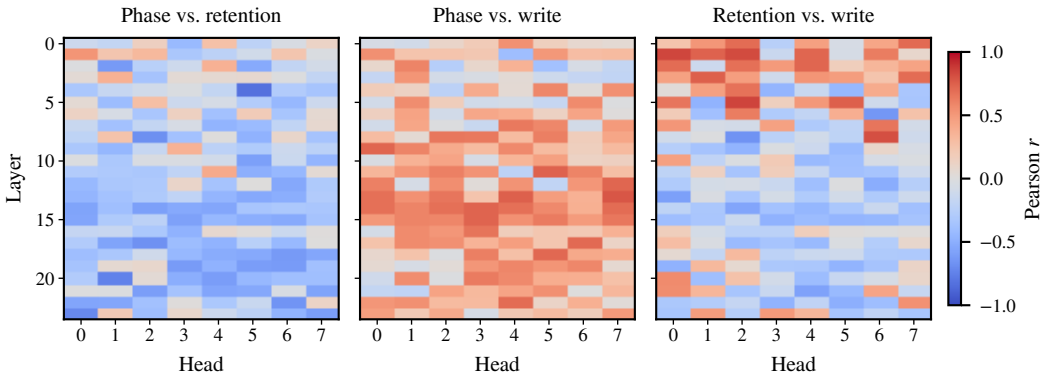

Figure 7: Token-wise coordination of the pretrained ODN-SO(2) controls on the same 128 WikiText-2 articles of 4,096 tokens used for the address measurements. Each cell is a layer–head Pearson correlation after within-article token centering.

F.1 Gate Coordination Across Evaluation Corpora

Table 2 reports the clock–write correlation across all evaluation corpora. Figure 7 shows the corresponding layer–head structure on the 128 WikiText-2 articles used for the pretrained-address analysis, together with clock–retention and retention–write correlations. Each gate is centered over tokens within an article, layer, and head before pooling across articles, so the maps describe token-level comodulation rather than differences in layer–head baselines.

F.2 Lag-Indexed Overwrite Recall by Lag

Figure 8 resolves the Lag-Indexed Overwrite Recall results in Table 3 by requested lag.

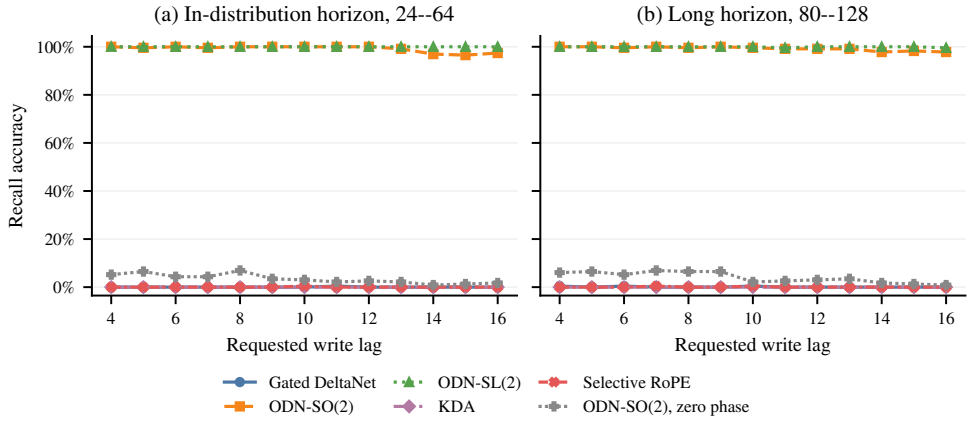

Figure 8: Lag-Indexed Overwrite Recall accuracy by requested write lag on (a) the in-distribution 24–64-write horizon and (b) the 80–128-write horizon.

F.3 Parameterization Ablations

Table 7 compares alternative parameterizations within each model family and repeats the default model as the reference. A tanh clock gate permits signed angular advance; independent clock and retention gates operate per oscillator; and the generic ODN-SL(2) variant parameterizes rotation, squeeze, and shear rates independently. The unit clock sets $\Delta_t = 1$ in the phase term only, so the retention term keeps its token-dependent control; the fixed spectrum freezes ω_r at initialization; and the odd-harmonic initialization sets $\omega_r = (2n_r + 1)\, 2\pi/T_{\max}$, with integers n_r spaced evenly from

0 to $\lfloor (T_{\max}/T_{\min} - 1)/2 \rfloor$, instead of the log-uniform spectrum of Equation 7, with frequencies still learnable.

Table 7: Parameterization comparisons at 340M parameters and 15B pretraining tokens. Default models are repeated from Table 4. Independent clock and retention gates operate per oscillator. Bold marks the best and underline the second-best result in each column within each model family, including ties.

Parameterization	Wiki. ppl↓	LMB. ppl↓	LMB. acc↑	PIQA acc↑	Hella. acc_n↑	Wino. acc↑	ARC-e acc↑	ARC-c acc_n↑	BoolQ acc↑	Avg. ↑
ODN-SO(2)										
Default (reference)	28.40	**32.46**	**34.02**	65.34	39.72	51.46	**59.22**	27.22	**60.70**	**48.24**
tanh clock gate	27.91	34.23	33.59	66.32	39.83	**52.25**	56.78	27.05	56.48	47.47
Independent retention gates	28.31	34.50	33.30	66.92	39.42	50.59	56.86	26.96	60.61	47.81
Independent clock gates	**27.69**	35.30	32.37	**67.41**	39.70	49.72	58.16	**29.69**	59.91	48.14
Unit clock	28.52	34.27	33.28	66.49	39.31	51.07	57.24	27.99	56.82	47.46
Fixed spectrum	28.23	34.81	32.78	67.36	39.57	51.30	57.37	26.96	60.15	47.93
Unit clock + fixed spectrum	28.97	36.20	32.33	66.38	**39.88**	50.59	57.32	28.84	48.38	46.25
Odd-harmonic spectrum init	28.13	36.37	32.10	66.92	39.24	51.78	57.49	27.13	60.43	47.87
ODN-SL(2)										
Default (reference)	28.02	**32.23**	33.67	66.97	39.53	**52.64**	**58.63**	27.73	60.89	**48.58**
Independent retention gates	28.36	35.41	32.27	**67.25**	**39.84**	50.83	58.08	28.24	59.02	47.93
Independent clock gates	**27.62**	33.19	32.33	66.38	39.45	52.41	58.12	**28.75**	59.82	48.18
Generic ODN-SL(2)	29.24	34.60	**34.31**	66.21	39.05	51.54	56.44	27.47	**61.10**	48.02

G Unified Chunkwise Parallel Algorithm

We derive one chunkwise algorithm for ODN-SO(2) and ODN-SL(2). Following the WY representation and UT transform used in Gated DeltaNet and KDA (Yang et al., 2025; Kimi Team, 2025), we first express the cumulative delta updates through auxiliary key and value vectors, then compute these vectors with a shared triangular solve. The derivation operates on the original fast-weight state and retains the full 2×2 transport of each oscillator.

The stacked state, per-oscillator transition, and complete recurrence are those of Equations 3–6 and Equation 2.

WY representation. Consider a chunk of b tokens, with local indices $1, \dots, b$ and incoming state S_{in}. Define the transport-only prefixes and interval products by

$$\Phi_0 = I, \qquad \Phi_i = T_i \cdots T_1, \qquad \Phi_{i \leftarrow j} = T_i \cdots T_{j+1}, \qquad \Phi_{i \leftarrow i} = I. \tag{35}$$

Let $\mathcal{P}_i$ be the cumulative transition including the delta erasures, and $\mathcal{H}_i$ its accumulated write term, so that $S_i = \mathcal{P}_i S_{\text{in}} + \mathcal{H}_i$. With $\mathcal{P}_0 = I$ and $\mathcal{H}_0 = 0$, their recurrences are

$$\mathcal{P}_i = (I - \beta_i k_i^\top k_i) T_i \mathcal{P}_{i-1}, \tag{36}$$

$$\mathcal{H}_i = (I - \beta_i k_i^\top k_i) T_i \mathcal{H}_{i-1} + \beta_i k_i^\top v_i. \tag{37}$$

They admit the following generalized WY representation:

$$\mathcal{P}_i = \Phi_i - \sum_{j=1}^{i} \Phi_{i \leftarrow j} k_j^\top w_j, \tag{38}$$

$$\mathcal{H}_i = \sum_{j=1}^{i} \Phi_{i \leftarrow j} k_j^\top u_j, \tag{39}$$

where the auxiliary rows $w_i \in \mathbb{R}^{1\times d}$ and $u_i \in \mathbb{R}^{1\times d_v}$ depend only on the chunk inputs:

$$w_i = \beta_i \left(k_i \Phi_i - \sum_{j<i} (\Psi_{kk})_{ij} w_j \right), \tag{40}$$

$$u_i = \beta_i \left(v_i - \sum_{j<i} (\Psi_{kk})_{ij} u_j \right), \qquad (\Psi_{kk})_{ij} = \begin{cases} k_i \Phi_{i\leftarrow j} k_j^\top, & j < i, \\ 0, & j \geq i. \end{cases} \tag{41}$$

These identities follow by induction: applying T_i transports all earlier terms to token i, and the current rank-one correction supplies the new w_i, u_i terms. The same argument holds for rotation and general $\mathrm{SL}(2,\mathbb{R})$ transport, including non-orthogonal flows.

UT transform. Stack w_i, u_i, v_i into W, U, V, and define the transported feature rows

$$\widehat{K}_{i,:} = k_i \Phi_i, \qquad \widehat{Q}_{i,:} = q_i \Phi_i, \qquad \mathcal{B} = \mathrm{Diag}(\beta_1, \ldots, \beta_b). \tag{42}$$

Equations 40–41 become $W = \mathcal{B}(\widehat{K} - \Psi_{kk} W)$ and $U = \mathcal{B}(V - \Psi_{kk} U)$. Both auxiliary matrices therefore share the same triangular factor:

$$M = (I + \mathcal{B}\Psi_{kk})^{-1}\mathcal{B}, \qquad W = M\widehat{K}, \qquad U = MV. \tag{43}$$

Here $M \in \mathbb{R}^{b\times b}$, $W \in \mathbb{R}^{b\times d}$, and $U \in \mathbb{R}^{b\times d_v}$. Since Ψ_{kk} is strictly lower triangular, $I + \mathcal{B}\Psi_{kk}$ has unit diagonal, including when a write gate is zero. The inverse denotes a triangular solve, implemented by forward substitution:

$$M_{i,:} = \beta_i \left(\mathbf{e}_i^\top - \sum_{j<i} (\Psi_{kk})_{ij} M_{j,:} \right), \tag{44}$$

where $\mathbf{e}_i$ is the ith standard basis vector. This is the same auxiliary-value convention as KDA: U is independent of the incoming state; the actual gated residual is $U - WS_{\mathrm{in}}$.

Chunkwise state update and output. Define the keys transported to the end of the chunk and the causal query–key interaction matrix by

$$\check{K}_{j,:} = (\Phi_{b\leftarrow j} k_j^\top)^\top, \qquad (\Psi_{qk})_{ij} = \begin{cases} q_i \Phi_{i\leftarrow j} k_j^\top, & j \leq i, \\ 0, & j > i. \end{cases} \tag{45}$$

Thus $\check{K} \in \mathbb{R}^{b\times d}$ and $\Psi_{qk} \in \mathbb{R}^{b\times b}$. In particular, the direction of transport in $\check{K}$ is specified before transposing; the flow need not be symmetric or orthogonal. Equations 38 and 39 give

$$\mathcal{P}_b = \Phi_b - \check{K}^\top W, \qquad \mathcal{H}_b = \check{K}^\top U. \tag{46}$$

Substitution yields the common chunk update for both models:

$$\Delta V = U - WS_{\mathrm{in}}, \tag{47}$$

$$S_{\mathrm{out}} = \Phi_b S_{\mathrm{in}} + \check{K}^\top \Delta V, \tag{48}$$

$$O = \widehat{Q} S_{\mathrm{in}} + \Psi_{qk} \Delta V. \tag{49}$$

The ith row of ΔV is exactly $\beta_i(v_i - k_i T_i S_{i-1})$. The strict mask in Ψ_{kk} evaluates this prediction before the current write, while the diagonal of Ψ_{qk} reads the state after that write. As in Gated DeltaNet, the kernel scales the readout by $d^{-1/2}$; this constant is omitted from Equations 1 and 2 and from the derivation above.

This follows the Gated DeltaNet/KDA pattern of a transported incoming state plus an auxiliary-value correction. Their scalar or diagonal interval retention factors are replaced here by $\Phi_{i\leftarrow j}$, while the shared M, W, U solve and the $U - WS_{\mathrm{in}}$ correction are retained. The expressions are algebraically exact for both parameterizations. Only the 2×2 transport blocks and the rectangular factors are stored; neither $\mathcal{P}_b$ nor $\check{K}^\top W$ is formed as a dense $d \times d$ matrix.

Algorithm 1 Unified chunkwise evaluation of ODN-SO(2) and ODN-SL(2)

Require: q, k, v, β, oscillator generators A_r, steps Δ_t, base log retentions $\log \rho_t$, guard multipliers $\log \kappa_{t,r}$, initial state S_0

Split each sequence into chunks of at most $b = 16$ tokens
for all chunks and heads in parallel **do**
    Scan $P_{i,r}, \lambda_{i,r}$ and form $\widehat{Q}, \widehat{K}, \Psi_{kk}, \Psi_{qk}, \check{K}, \Phi_b$
    Solve $(I + \mathcal{B}\Psi_{kk})M = \mathcal{B}$ and set $W \leftarrow M\widehat{K}$
end for
for all sequences, heads, and value-column tiles in parallel **do**
    $S \leftarrow S_0$ for the assigned head and value columns
    for chunks in sequence order **do**
        $U \leftarrow MV; \quad \Delta V \leftarrow U - WS$
        $O \leftarrow d^{-1/2}(\widehat{Q}S + \Psi_{qk}\Delta V)$
        $S \leftarrow \Phi_b S + \check{K}^\top \Delta V$
    end for
end for
return Concatenated chunk outputs and final states

Computing the factors with separate decay scans. The shared kernel construction scans the determinant-one flows and their log decays separately. For each oscillator r, define

$$P_{i,r} = \exp(\Delta_i A_r) \cdots \exp(\Delta_1 A_r), \qquad \lambda_{i,r} = \sum_{a=1}^{i} (\log \rho_a + \log \kappa_{a,r}) . \tag{50}$$

An associative scan builds $P_{i,r}$, and a prefix sum builds $\lambda_{i,r}$. All features and interactions above use the same per-oscillator factors

$$\Phi_{i,r} = e^{\lambda_{i,r}} P_{i,r}, \qquad \Phi_{i \leftarrow j,r} = e^{\lambda_{i,r} - \lambda_{j,r}} P_{i,r} P_{j,r}^{-1}. \tag{51}$$

For example, if $k_{i,r}$ denotes the two key components of oscillator r,

$$(\Psi_{kk})_{ij} = \sum_{r=1}^{m} e^{\lambda_{i,r} - \lambda_{j,r}} \, (k_{i,r} P_{i,r})(P_{j,r}^{-1} k_{j,r}^\top), \qquad j < i. \tag{52}$$

Replacing the target key by the query gives the entries of Ψ_{qk} with $j \le i$. The source vector is obtained by a 2×2 solve against $P_{j,r}$, which is valid for both ODN-SO(2) and ODN-SL(2). Decay enters only through prefixes and causal interval factors, whose exponents are non-positive; the implementation never inverts the decay. Prefixes restart at each chunk boundary.

For ODN-SO(2), $\lambda_{i,r}$ is shared across oscillators, so the interval retention factor multiplies the result of the query–key matrix products. For guarded ODN-SL(2), the oscillator-dependent decay stays inside the sum in Equation 52; the kernel accumulates interaction columns over oscillators. These are two implementations of the same factor construction, followed by identical triangular solves and state updates.

Implementation and complexity. Algorithm 1 follows the kernel schedule: factor construction is independent across chunks, while the state advances in chunk order. The kernel forms $U = MV$ and $\Delta V = U - WS_{\text{in}}$ during state application.

The kernels use a fixed internal chunk width of $b = 16$ tokens. Packed sequences have separate chunk indices, and partial final chunks mask padded writes and use identity transitions. Scans, triangular factors, residuals, and recurrent accumulators use FP32; stored feature factors use BF16, and matrix products use Triton's TF32 path. For sequence length L and chunk width b, forward arithmetic per head is

$$O\big(Ldd_v + Lb(d + d_v) + Lb^2\big) , \tag{53}$$

including $O(b^3)$ work for the triangular factor per chunk. For fixed b, d, d_v, work and training storage are linear in L. The state dependency spans $\lceil L/b \rceil$ chunk steps, and inference carries the $d \times d_v$ memory between chunks.

H Experimental Details

H.1 Language Models

Architecture. Every 340M model uses 24 layers, hidden size 1,024, 8 heads with head dimension 128 for keys, queries, and values, a 32K-token vocabulary (the `fla-hub/delta_net-1.3B-100B` tokenizer), untied embeddings, and the short convolution and gated RMSNorm output of Gated DeltaNet. Gated DeltaNet, KDA, and Selective RoPE use their released layer implementations with the same backbone; Selective RoPE is applied on the Gated DeltaNet layer with the full-rank phase projection. ODN-SO(2) and ODN-SL(2) use the chunkwise kernel of Appendix G.

ODN gates and initialization. Each gate is computed from a learned linear projection of x_t. The write gate is $\beta_t = \sigma(\cdot)$ without further scaling. The clock step is $\Delta_t = \sigma(\cdot)$ with zero-initialized weights and bias, so $\Delta_t = 0.5$ at initialization. Retention follows Gated DeltaNet, $\log \rho_t = -\nu \operatorname{softplus}(\cdot + b_\rho)$, with a per-head rate $\nu \sim \mathcal{U}(0, 16)$ and b_ρ chosen so that $\operatorname{softplus}(b_\rho)$ is log-uniform on $[10^{-3}, 10^{-1}]$. The periods $2\pi/\omega_r$ start at the midpoints of m equal bins in log period over $[T_{\min}, T_{\max}] = [256, 16{,}384]$, a stratified form of Equation 7 with $\omega_{\min} = 2\pi/T_{\max}$ and $\omega_{\max} = 2\pi/T_{\min}$. Each $\log \omega_r$ is perturbed by Gaussian noise of standard deviation 0.05, and the spectrum is cyclically shifted across heads. The frequencies used in the transport pass through a soft cap,

$$\log \omega_r \;\leftarrow\; \log \omega_{\text{cap}} - \operatorname{softplus}(\log \omega_{\text{cap}} - \log \omega_r)\,, \qquad \omega_{\text{cap}} = 0.99\pi, \tag{54}$$

so that a single step $\omega_r \Delta_t$ never reaches π. ODN-SL(2) generators start at $\phi_r = \psi_r = 0$.

Pretraining. All models are trained on the `sample-100BT` split of FineWeb-Edu with AdamW ($\epsilon = 10^{-15}$, weight decay 0.01), peak learning rate 3×10^{-4}, 1,024 warmup steps, cosine decay to 10% of the peak, gradient clipping at norm 1.0, and seed 42. Each optimizer step processes $8 \times 65{,}536$ tokens packed as variable-length documents into 4,096-token contexts, for 30,720 steps. Frequencies, retention parameters, and transport generators are excluded from weight decay. Training uses BF16 compute with FP32 master parameters and optimizer state.

Zero-shot evaluation. WikiText-2 and LAMBADA perplexity and the accuracies in Tables 4–7 are computed with the `lm-evaluation-harness` in zero-shot mode with a 2,048-token maximum length. HellaSwag and ARC-Challenge report normalized accuracy.

H.2 Mechanism Analyses

Controlled continuous-write experiment. This experiment isolates the similar-key interference analyzed in Section 4: a target association is followed by weak isotropic filler writes and by periodic strong writes whose keys resemble the target key. Each memory has a 128×32 state ($m = 64$ oscillators) and unit retention. Oscillation-based content addressing uses a unit clock and a fixed spectrum at the midpoints of 64 log-period bins over 256–16,384 tokens, without the initialization noise and with the soft cap above, giving periods of 266.47–15,862.24 tokens; static content addressing uses no clock advance.

The target key k_0 has equal coordinates, and the target value v_0 is a fixed 32-dimensional unit-norm Rademacher vector. After 128 isotropic prefill writes, the target is written at position zero with unit strength, and each of the next 32,768 positions contains one write. Every 1,024th position is a unit-strength event with key $s\,k_0 + \sqrt{1 - s^2}\,z$, where z is a normalized Gaussian direction orthogonal to k_0 and $s = 0.9$ is the cosine to the target key; the remaining 32,736 positions are isotropic filler writes. Prefill and filler writes use $\beta_{\text{fill}} = 0.0025$. Reads occur after the write at each reported position and do not modify memory. Oscillation-based content addressing queries with k_0 transported by its age, and static content addressing queries with k_0. Figure 1(e) reports the recall error $\|o_t - v_0\|_2^2$ of the full read o_t, including the filler background, with $\|v_0\|_2 = 1$. Curves show the mean over 64 matched noise realizations, and shading spans the 10th–90th percentiles.

Pretrained address measurements. We use 128 WikiText-2 articles, each truncated to its first 4,096 tokens, and drop the leading beginning-of-sequence position, leaving $L = 4{,}095$ positions

per article. Every statistic is computed separately for each of the $24 \times 8 = 192$ heads, averaged over articles for each head, and then averaged over heads. Within one head, the content and phase addresses of position i are the unit row vectors

$$x_i^{\text{content}} = \hat{k}_i, \qquad x_i^{\text{phase}} = \hat{k}_i \, \mathcal{R}(H_i), \qquad G_{ij} = x_i x_j^\top, \tag{55}$$

where $\hat{k}_i = k_i / \|k_i\|_2$. This phase address differs from the read-frame address of Equation 19 by the common rotation $\mathcal{R}(H_t)^\top$, which leaves G and every statistic below unchanged. Gated DeltaNet has only content addresses.

Mean squared overlap averages G_{ij}^2 over position pairs,

$$\overline{G^2} = \frac{1}{|\mathcal{I}|} \sum_{(i,j)\in\mathcal{I}} G_{ij}^2, \tag{56}$$

where $\mathcal{I}$ holds 4,096 pairs (i, j) with $i \neq j$, sampled uniformly with replacement per article. *Same-token squared overlap* restricts the pairs to repeated occurrences of one token. Let $\mathcal{V}$ be the token types that occur at least eight times in the article, capped at 24 types, and let O_v be up to 64 occurrence positions of type v. Pooling pairs over types,

$$\overline{G^2}_{\text{same}} = \frac{\sum_{v\in\mathcal{V}} \sum_{i<j,\; i,j\in O_v} G_{ij}^2}{\sum_{v\in\mathcal{V}} \binom{|O_v|}{2}}. \tag{57}$$

Participation ratio is computed from the eigenvalues $\lambda_1 \geq \cdots \geq \lambda_d$ of the address second-moment matrix over all L positions,

$$\Gamma = \frac{1}{L} \sum_{i=1}^{L} x_i^\top x_i, \qquad \text{PR} = \frac{\left(\sum_{n=1}^{d} \lambda_n\right)^2}{\sum_{n=1}^{d} \lambda_n^2} = \frac{1}{\sum_{n=1}^{d} \lambda_n^2}. \tag{58}$$

Because every address has unit norm, $\sum_n \lambda_n = \operatorname{tr}\Gamma = 1$, so λ_n is the share of address variance along the nth principal direction. Moreover, $\sum_n \lambda_n^2 = \operatorname{tr}\Gamma^2$ is the mean of G_{ij}^2 over all position pairs, including $i = j$, so the participation ratio is the reciprocal of a squared overlap. Figure 2 uses the same spectrum: panel (a) plots the cumulative explained variance $E(n) = \sum_{n'\leq n} \lambda_{n'}$ against n/d, and panel (b) plots $n_{0.9}/d$ with $n_{0.9} = \min\{n : E(n) \geq 0.9\}$. Lines show the median over heads and bands the interquartile range.

Shared-component decomposition. For each token type $v \in \mathcal{V}$, we estimate the mean key $\bar{k}$ as the mean of the unit content keys at O_v, set $c = \|\bar{k}\|_2^2$ and $u = \bar{k}/\sqrt{c}$, and transport $\bar{k}$ with the learned clock to obtain the shared term $c\, C_u(H_i - H_j)$ of Proposition 2. We compare the measured phase overlap with its shared part $c^2 C_u^2$ and with a permutation baseline. The baseline permutes the content keys among the occurrences of the token while each position keeps its own transport; this preserves $\bar{k}$ and Σ but removes any dependence of a key on its write time, which is present in the model because the clock and the keys are computed from the same inputs. We average eight permutations. Estimating $\bar{k}$ from $|O_v|$ occurrences raises c in expectation by $(1 - c)/|O_v|$, so the shared terms reported here and in Table 1 are upper estimates. Pairs are those of the same-token overlap.

The ODN-SO(2) content overlap is 0.411, of which the shared term c^2 is 0.381. The phase overlap is 0.147, of which the shared term $c^2 C_u^2$ is 0.126; the permutation baseline gives 0.144, about 2% below the measurement, and the cross term $2\,\mathbb{E}[c\, C_u(G_{ij} - c\, C_u)]$ is below 10^{-3}. The overall agreement is dominated by distant pairs. By gap, the ratio of measured to permuted overlap is 1.19 at one token, 0.94–1.03 at two to seven tokens, 1.04–1.11 from 8 to 255 tokens, and within 3% of one beyond: occurrences within a few hundred tokens share context, so their keys are more alike than independent draws. The pair mean of C_u^2 is 0.31; by gap it exceeds 0.97 below 32 tokens, is 0.56 at 128–255 tokens, and falls to 0.17 at 2K–4K tokens.

The reference curve of Figure 2c is the root mean square of C_u over log-uniform frequency draws with equal oscillator energies, $\left(\bar{C}(a)^2 + m^{-1}[(1 + \bar{C}(2a))/2 - \bar{C}(a)^2]\right)^{1/2}$, evaluated at $a = \bar{\Delta}_\ell g$ for token gap g and each layer's mean clock rate $\bar{\Delta}_\ell$, and averaged over layers. It matches the

measurement up to a few hundred tokens and falls below it at longer gaps, which suggests that shared key components concentrate their energy on slow oscillators or that training lowered the frequencies. The learned clock advances by 0.45 per token on average; a unit clock would move the same kernel to shorter token gaps and lower the phase overlap to 0.08.

H.3 Synthetic Tasks

MQAR. We use the mixed-length Zoology generator with vocabulary 8,192. Training data contain six segments of (sequence length, key–value associations): $(128, 8)$, $(256, 16)$, $(256, 32)$, $(256, 64)$, $(512, 128)$, and $(1024, 256)$, with 20,000 examples each; the test segments are $(128, 32)$, $(256, 64)$, $(512, 128)$, and $(1024, 256)$. Every model has two layers, width 128, two heads of dimension 64, and no short convolution, and is trained for 50 epochs with AdamW (learning rate 10^{-3}, weight decay 0.1), batch size 256, and gradient clipping at 1.0.

Lag-Indexed Overwrite Recall. Each write is serialized as `WRITE key value`, and the stream ends with `QUERY key DELAY_D RECALL`. All writes in an example use the same key, values are sampled without replacement, and the supervised target is the value written exactly D write events before the query. Training contains 30,000 examples with 24–64 writes and D sampled uniformly from $[4, 16]$. Evaluation uses 3,000 examples over the same horizon and 3,000 over 80–128 writes, stratified by exact lag. Models have two layers, width 128, two heads, vocabulary size 8,192, and no learned absolute position embedding or short convolution. We train for 50 epochs with AdamW (learning rate 10^{-3}, weight decay 0.1), batch size 128, evaluation batch size 256, and gradient clipping at 1.0. ODN initializes its periods over $[T_{\min}, T_{\max}] = [4, 256]$, and the no-clock-advance ablation of Table 3 sets $\Delta_t = 0$.

S2 state tracking. The task follows the state-tracking setup of Selective RoPE (Movahedi et al., 2026): one layer of width 384 with 12 heads of dimension 32 for every model, 2,000,000 training sequences of length 128 with dense prefix supervision, batch size 1,024, AdamW (learning rate 10^{-3}, weight decay 10^{-6}) with 1,000 warmup steps and cosine decay to 10^{-5}, gradient clipping at 1.0, and seeds 555, 666, 777, 888, and 999. Prefix-exact accuracy counts a sequence as correct only when every running product is correct. Selective RoPE uses the released state-tracking configuration (rank-32 phase projection, frozen temperature).

A3 state tracking. We use the same data sizes, sequence lengths, optimizer schedule, width, heads, dense prefix supervision, and five seeds as for S2, but replace the two-element group by A3 and use two model layers. Models are trained for 10,000 optimizer steps; Figure 3c shows the length-128 evaluation recorded every 100 steps.