Semidirect Fourier Delta Attention

Phase-Controlled Delta Memory with Constructive Chunk-WY Kernels

Paper by Tiantian Zhang (Crystal)
Columbia University

arXiv:2607.11897v1
Research discussion — 20 October 2026

Phase memory + delta updates + exact chunk-local algebra

Can phase memory preserve a useful chunk structure?

The design problem

Add cyclic state dynamics to KDA while keeping structured products of recurrent transitions.

1. **Architecture:** replace real decay with phase–decay.
2. **Algebra:** explicitly construct a chunk-WY factorization.
3. **Evidence:** exact identities and small counter experiments.

Fused GPU performance and language-model gains remain open.

Paper, Secs. 1, 4, 7–8, 10; PDF pp. 2–3, 7–10, 16–22.

Rotate the memory, then correct its prediction

$$\boxed{S_t = (I - \beta_t k_t k_t^*)\Lambda_t S_{t-1} + \beta_t k_t v_t^*} \qquad \Lambda_t = \mathrm{diag}(\alpha_t \odot e^{i\theta_t})$$

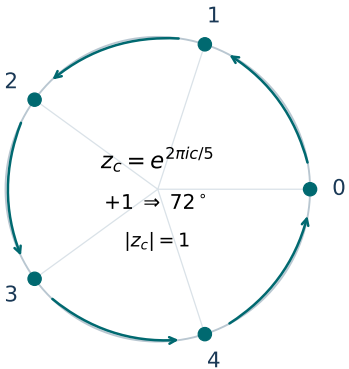
$$\begin{aligned}
\bar{S}_t &= \Lambda_t S_{t-1} && \text{transport and decay,} \\
\widehat{v}_t &= \bar{S}_t^* k_t && \text{read the current prediction,} \\
S_t &= \bar{S}_t + \beta_t k_t (v_t - \widehat{v}_t)^* && \text{write the residual.}
\end{aligned}$$

$S_t \in \mathbb{C}^{d \times d_v}$; $\quad q_t, k_t \in \mathbb{C}^d$; $\quad v_t \in \mathbb{C}^{d_v}$; $\quad o_t = S_t^* q_t$.

The same key erases and writes. Here * is conjugate transpose.

Paper, Sec. 3, Eqs. (13)–(19), PDF p. 5. Three-step form: explanatory rewrite.

A phase coordinate is a bounded modular counter



With $z_0 = 1$, $\alpha_t = 1$, $\beta_t = 0$:

$$z_t = e^{2\pi i a_t/m} z_{t-1}.$$

Hence

$$z_t = e^{2\pi i \sum_{s \le t} a_s/m}, \quad |z_t| = 1.$$

- Decode the nearest phase prototype.
- A zero initial state cannot rotate into a counter.
- No FFT is required.

Paper, Prop. 3.4, PDF p. 7; App. G.2, p. 38. Diagram: reader-created teaching illustration.

Different control families support different claims

Family	Transition and scope
Real KDA subfamily	$(I - \beta kk^\top)\,\mathrm{diag}(\alpha)$; real decay
Default complex SFDA	$(I - \beta kk^*)\,\mathrm{diag}(\alpha e^{i\theta})$; phase memory
Structured generalization	$\Lambda - ur^*$; cheap reflections or shifts may be added
Unrestricted template	$(I - \beta kk^*)\widetilde{A}$; arbitrary symbol-conditioned $\widetilde{A}$

One-hot finite-state realizability uses the last row. Reflections and stack shifts require their corresponding controls.

Paper, Table 2, PDF p. 6; Thm. 4.7, p. 9; Sec. 5, pp. 11–13.

WY factors retain interactions between delta updates

Set $u_t = \beta_t k_t$ and $r_t = \Lambda_t^* k_t$, so

$$A_t = \Lambda_t - u_t r_t^*.$$

For two tokens,

$$A_2 A_1 = \Lambda_2 \Lambda_1 - \Lambda_2 u_1 r_1^* - u_2 r_2^* \Lambda_1 + \underbrace{u_2 (r_2^* u_1) r_1^*}_{\text{interaction of the two corrections}}.$$

What the factors store

Γ: structured control product; Y, W: correction directions;
M: interactions between corrections.

r_t is the right WY vector, *not* the attention value.

Paper, Eq. (26), Thms. 4.1–4.2, PDF pp. 7–8. Two-step expansion is an explanatory derivation.

The chunk factorization is constructive

For $P_t = A_t \cdots A_1$, initialize $\Gamma_0 = I$ and empty Y_0, W_0, M_0:

$$\begin{aligned}
\Gamma_t &= \Lambda_t \Gamma_{t-1}, \\
Y_t &= [\, \Lambda_t Y_{t-1},\ u_t \,], \\
W_t &= [\, W_{t-1},\ \Gamma_{t-1}^* r_t \,], \\
M_t &= \begin{bmatrix} M_{t-1} & 0 \\ -r_t^* Y_{t-1} M_{t-1} & 1 \end{bmatrix}.
\end{aligned}$$

$$\boxed{P_t = \Gamma_t - Y_t M_t W_t^*}$$

$Y_t, W_t \in \mathbb{C}^{d \times t}$, $M_t \in \mathbb{C}^{t \times t}$. No inverse of Λ_t is required.

Paper, Thm. 4.2, Eqs. (34)–(39), PDF p. 8; proof in App. A.4, p. 25.

A complete chunk summary must include new writes

With $B_t = \beta_t k_t v_t^*$,

$$B_C^{\text{chunk}} = \sum_{i=1}^{C} A_C \cdots A_{i+1} B_i.$$

Compute it by the same recurrence starting from zero:

$$\bar{S}_0 = 0, \qquad \bar{S}_t = A_t \bar{S}_{t-1} + B_t.$$

Then apply the summary to the incoming state:

$$\boxed{S_{\text{out}} = \Gamma_C S_{\text{in}} - Y_C\big(M_C(W_C^* S_{\text{in}})\big) + B_C^{\text{chunk}}}.$$

- Each write is transported by all subsequent transitions.
- Prefix factors also recover every intra-chunk token output.

Paper, Thms. 4.4–4.5, PDF pp. 8–9; Algorithm 1, p. 14.

Keep the rank budget local to each chunk

Exact local representation

$$\operatorname{rank}(P_C - \Gamma_C) \le \min(C, d).$$

1. Build factors inside a chunk.
2. Apply them to boundary state S.
3. Start new factors in the next chunk.

Why rank can grow

$$A_i = I - e_i e_i^*$$

$$A_r \cdots A_1 = I - \sum_{i=1}^{r} e_i e_i^*.$$

The correction has rank r for $r \le d$.

A globally composed exact summary can grow to rank d. Small-rank recompression generally introduces approximation.

Paper, Thm. 5.3 and Prop. 5.4, PDF pp. 11–12; Sec. 6, p. 15.

Reference costs expose the remaining systems work

Per-chunk operation	Reference cost
Factor storage + write summary	$O(d + dC + C^2 + dd_v)$
Explicit factor construction	$O(dC^2 + C^3)$
Zero-state write recurrence	$O(Cdd_v)$
Apply to the boundary state	$O(dd_v + dCd_v + C^2 d_v)$

- Fixed C gives linear scaling in sequence length T.
- If $d_v = d$, the state update still includes a d^2 term.
- These bounds do not measure all outputs, backward passes, device traffic, or wall-clock throughput.

Paper, Thm. 6.2, PDF p. 15; App. B.6 and C, pp. 27–30. Width-scaling interpretation added by the reader.

Nonexpansive transitions can still accumulate writes

If $\|\Lambda_t\|_2 \le 1$ and $0 \le \beta_t \|k_t\|_2^2 \le 2$, then

$$\|A_t\|_2 \le 1, \qquad \|S_t\|_F \le \|S_0\|_F + \sum_{i=1}^{t} \|B_i\|_F .$$

Proof idea

$I - \beta k k^*$ has eigenvalues $1 - \beta \|k\|^2$ along k and 1 on its orthogonal complement.

- Uniform decay $\|\Lambda_t\| \le \rho < 1$ gives a geometric bound on accumulated writes.
- Stable state transitions do not automatically imply well-conditioned intermediate WY factors.

Paper, Props. 6.3–6.4, Eqs. (74)–(77), PDF pp. 15–16.

Compact memories require specific control operators

Memory	State size	Required mechanism
Product cyclic counters	ℓ complex coordinates	Phase-only; no decay or write
Dihedral D_m	2×2 real matrix	Rotations *and reflections*
Registers / resets	$d \times d_v$	Erase row i, then write row i
Stack of depth h	$h \times d_v$	Zero-fill shift matrices

Registers already fit KDA. Reflection and shift theorems extend the default phase control family.

Paper, Thms. 5.5, 5.7, 5.9, 5.11, PDF pp. 12–13; App. G, pp. 37–38.

Routing a write also requires protecting old memory

Desired update: $x^+ = G_\sigma x + w_\sigma r_\sigma$.

$$G_\sigma = R_\sigma \widetilde{G}_\sigma, \qquad w_\sigma = R_\sigma k, \qquad k^\top \widetilde{G}_\sigma x = 0.$$

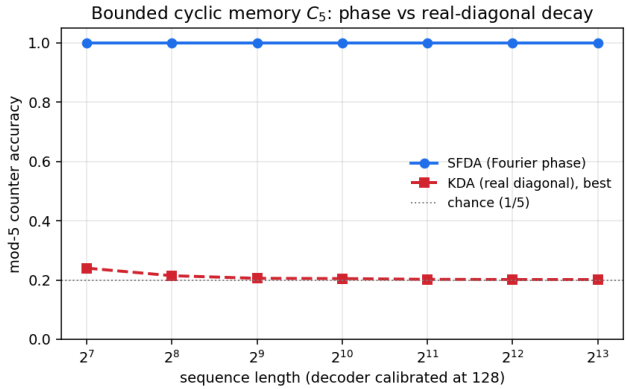
Under these conditions,

$$R_\sigma \left[(I - kk^\top) \widetilde{G}_\sigma x + k r_\sigma \right] = G_\sigma x + w_\sigma r_\sigma.$$

- w_σ must lie in the allowed control orbit of k.
- The orthogonality condition neutralizes destructive erasure.
- Fixed block rotations cannot move a key between blocks.

Paper, Sec. 4.2, PDF p. 10; Thm. F.2 and Prop. F.5, pp. 35–36.

Constructed phase memory stays exact to length 8192



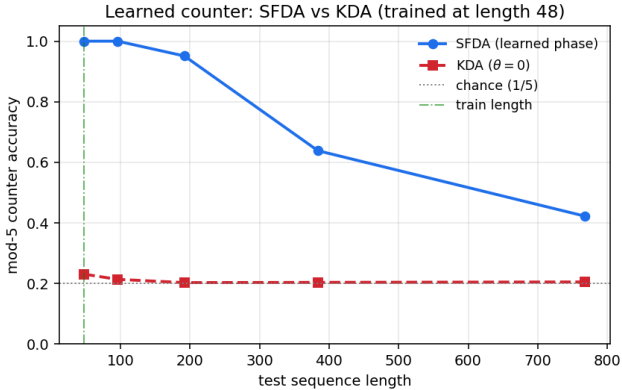
Constructed mod-5 counter

- One real 2D phase block.
- Decoder calibrated at 128.
- At 8192: 1.000 vs. 0.201.

The baseline is a low-dimensional real-diagonal control model.

Paper, Sec. 7, Fig. 1(a), Table 3, PDF pp. 16–17; App. H.2, p. 40. Original author figure.

Learned phase calibration limits extrapolation



Train at length 48.

Length	SFDA
96	1.000
192	0.951
384	0.638
768	0.422

A persistent per-step phase error can accumulate with length.

Paper, Fig. 1(b), Table 4, PDF pp. 17–18; App. H.3, p. 40. Original author figure; error accumulation is the author's interpretation.

The mod-3 study uses a separate evaluation protocol

Model and training

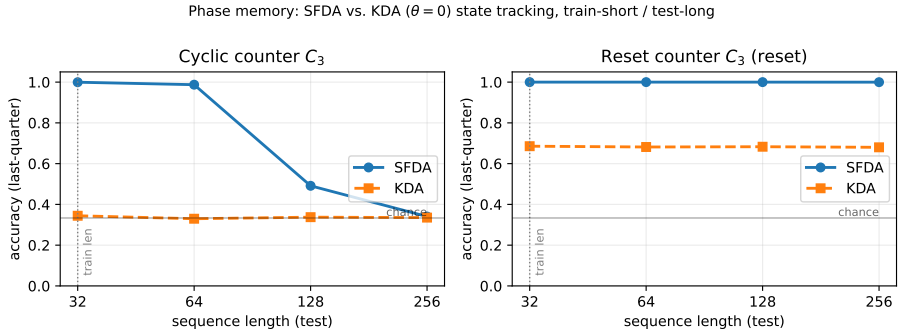
- One layer, one head.
- State width $d = 16$.
- Linear readout; ~ 8.3k parameters.
- AdamW + cosine schedule.
- Best in-distribution validation checkpoint.

Comparison and measurement

- Same architecture and parameter count.
- KDA: force $\theta_t = 0$.
- Train at $T = 32$; test to 256.
- Accuracy on last-quarter positions.
- Mean over two seeds; chance $1/3$.

Paper, Sec. 8.2, PDF pp. 18–19. These settings are not assigned to the mod-5 study.

Resets preserve accuracy; uninterrupted counting degrades



Cyclic C_3 at $T = 256$
SFDA 0.341 vs. KDA 0.335.
Nearly back to chance after 8× extrapolation.

Reset C_3 at $T = 256$
SFDA 1.000 vs. KDA 0.680.
Resets shorten the effective memory horizon.

Paper, Table 6, PDF p. 19; Fig. 2, p. 20. Original author figure. Reset interval distribution is not detailed.

Numerical checks support the algebraic identities

Identity / property	Table 5	Table 8
Constructive chunk-WY	5.7×10^{-10}	1.9×10^{-15}
Affine chunk transfer	9.5×10^{-8}	2.4×10^{-16}
KDA reduction at $\theta = 0$	0	0
Stability $\|A_t\|_2 \leq 1$	2000 samples	max 0.9977

- Table 5 reports operator/state residuals.
- Table 8 reports worst-case relative errors over its checks.
- The two test suites are not specified on a common precision and normalization basis.

Paper, Sec. 8.1, Table 5, PDF p. 18; App. H.1, Table 8, pp. 39–40. All values above are author-reported.

Direction reversal and systems scaling remain unresolved

A concrete failure case

On the dihedral D_3 task, SFDA reaches only about 0.47 in-distribution. A full length-extrapolation table is not provided.

- The exact dihedral construction requires reflection control; the experimental implementation of this extension is unclear.
- No fused Triton/CUDA throughput or language-model results.
- Ratios 7:1 and 15:1 versus KDA 3:1 are future targets.
- The source package lists experiment script names but does not include those scripts.

Paper, Sec. 8.2, PDF p. 19; Sec. 10, pp. 21–22; App. E and H.4, pp. 33–34, 40–41. Source-package observation by the reader.

The useful contribution couples dynamics to computation

1. **Phase supplies periodic dynamics.**
   Delta updates supply addressable residual writes.
2. **Constructive WY makes the chunk algebra usable.**
   Preserve the write summary and the chunk-local rank budget.
3. **Learned accuracy depends on phase precision.**
   Reset and uninterrupted tasks probe different memory horizons.

The next decisive tests

Match real state bytes, isolate reflection control, measure phase drift, and benchmark complete fused forward/backward kernels.

Synthesis of Secs. 3–8 and 10. Proposed next tests are reader analysis.

Paper and discussion guide

Tiantian Zhang (Crystal).
Semidirect Fourier Delta Attention: Phase-Controlled Delta Memory with Constructive Chunk-WY Kernels.
arXiv:2607.11897v1.

arxiv.org/abs/2607.11897v1

- Main update and exact recursion: pp. 5–9.
- Compact controls, cost, stability: pp. 11–16.
- Reported counter results: pp. 17–20.
- Routing theorem: pp. 34–37.

All locators refer to the 41-page v1 PDF. Paper and author figures: CC BY 4.0. The phase-circle diagram is a reader-created teaching illustration.

Backup: block-WY closure accounts for cross terms

For $A_i = \Gamma_i - Y_i M_i W_i^*$, write

$$A_2 A_1 = \Gamma_{21} - Y_{21} M_{21} W_{21}^*,$$

with

$$\begin{aligned}
\Gamma_{21} &= \Gamma_2 \Gamma_1, \\
Y_{21} &= [\, \Gamma_2 Y_1, \;\; Y_2 \,], \\
W_{21} &= [\, W_1, \;\; \Gamma_1^* W_2 \,], \\
M_{21} &= \begin{bmatrix} M_1 & 0 \\ -M_2 W_2^* Y_1 M_1 & M_2 \end{bmatrix}.
\end{aligned}$$

The represented rank budget grows from (r_1, r_2) to $r_1 + r_2$.

Paper, Thm. 4.1, Eqs. (27)–(33), PDF p. 7.

Backup: prefix factors produce every token output

For a prefix $r \le C$,

$$P_r = \Gamma_r - Y_r M_r W_r^*, \qquad S_r = P_r S_0 + B_r^{\mathrm{chunk}}.$$

Therefore

$$o_r = S_0^* P_r^* q_r + (B_r^{\mathrm{chunk}})^* q_r,$$

where

$$P_r^* q_r = \Gamma_r^* q_r - W_r M_r^* Y_r^* q_r.$$

- Both the carried state and local writes contribute.
- This establishes exact outputs.
- An efficient batched implementation must also manage prefix storage, repeated work, and backward computation.

Paper, Thm. 4.5, PDF p. 9. Adjoint expansion and implementation implications are reader analysis.

Backup: realifying complex rank one gives real rank two

For a complex key, define

$$a = \begin{bmatrix} \Re k \\ \Im k \end{bmatrix}, \qquad J = \begin{bmatrix} 0 & -I \\ I & 0 \end{bmatrix}.$$

The complete operator satisfies

$$\mathcal{R}(kk^*) = aa^\top + (Ja)(Ja)^\top.$$

Example: $k = 1 \in \mathbb{C}$ gives

$$\mathcal{R}(kk^*) = I_2 \quad \text{but} \quad aa^\top = \operatorname{diag}(1, 0).$$

Complex SFDA needs a complex inner product or paired real updates. A real rank-one delta plus block rotations is a distinct valid WY model.

Reader derivation clarifying the whole-update equivalence in Eqs. (2)–(3). Paper Prop. 3.3 and Lemma A.1 establish phase-control realification, PDF pp. 6, 24.

Backup: affine lifting uses one extra coordinate

For any affine system $x^+ = A_\sigma x + b_\sigma$, use

$$\phi(x) = \begin{bmatrix} x \\ 1 \end{bmatrix}, \qquad \widetilde{A}_\sigma = \begin{bmatrix} A_\sigma & b_\sigma \\ 0 & 1 \end{bmatrix}, \qquad \beta_\sigma = 0.$$

Then

$$\widetilde{A}_\sigma \phi(x) = \phi(A_\sigma x + b_\sigma), \qquad D = d + 1.$$

Scope of the open problem

The unrestricted template in Problem D.1 admits this standard lift. The interesting question needs structured-control, stability, or microstep constraints.

Reader analysis of App. D.1 and D.5, PDF pp. 30–33. This construction is not a phase-only SFDA kernel.

Backup: exact counter constructions use exact endpoints

Representation conditions

- Phase counter: $\beta = 0$, $\alpha = 1$.
- Register replacement: $\beta = 1$.
- Finite-logit sigmoid gates generally approximate these endpoints.

A phase-error illustration

$$\Delta\theta_T = T\varepsilon$$

for a constant per-step error ε.

$$|\Delta\theta_T| < \frac{\pi}{m}$$

keeps nearest-prototype decoding inside the correct angular cell.

Exact expressivity, numerical precision, and learnability are separate parts of the design.

Paper, Eqs. (15)–(18), Prop. 3.4, Thm. 5.9, PDF pp. 5, 7, 13. Endpoint and phase-error discussion: reader analysis, not a fitted error model.