Complex KDA

Understanding and Enhancing the Expressivity of Kimi Delta Attention

Paper by Julien Siems, Riccardo Grazzi,
Korbinian Pöppel, and collaborators

arXiv:2609.24797v1 | Paper discussion, 8 October 2026

Signed gates turn one delta update into a rotation.

State tracking needs ordered transformations

The task

Read group elements $g_1, \ldots, g_t$ and recover every prefix product:

$$s_t = g_t \cdots g_1.$$

- Parity needs a sign flip; modular counting needs a cycle.
- Permutation tracking needs noncommuting transformations.
- Inputs range over all group elements, not only swaps.

A fixed-size state is useful only if its update can represent the required operations.

Paper, Secs. 1–2 and 5, PDF pp. 1–3 and 6; App. B.1, p. 26.

CKDA keeps the original gate–delta–readout structure

$$H_t = (I - \beta_t k_t k_t^\top)\,\mathrm{Diag}(\alpha_t) H_{t-1} + \beta_t k_t v_t^\top,$$
$$o_t = H_t^\top q_t, \qquad \|k_t\|_2 = 1.$$

$H_t \in \mathbb{R}^{n \times d_v}$: matrix state

$k_t, q_t \in \mathbb{R}^n$, $v_t \in \mathbb{R}^{d_v}$

$\alpha_t \in \mathbb{R}^n$: channel gates

1. Gate the old state.
2. Correct the key–value prediction.
3. Query the updated state.

CKDA changes the ranges to $\alpha_t \in [-1, 1]^n$, $\beta_t \in [0, 2]$.

Paper, Eq. (1), PDF p. 3; Sec. 3, p. 4; App. B.2, pp. 26–27.

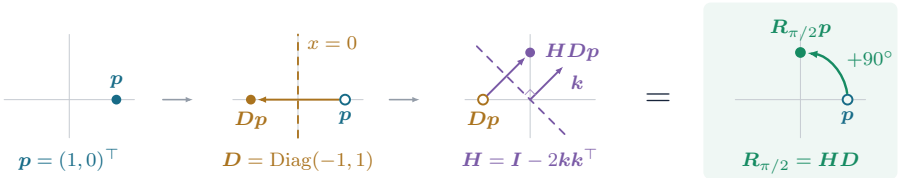
Both range extensions are needed for complex eigenvalues

Channel gates	Update rate	Non-real eigenvalues possible?
$[0, 1]$	$[0, 1]$	No
$[0, 1]$	$[0, 2]$	No
$[-1, 1]$	$[0, 1]$	No
$[-1, 1]$	$[0, 2]$	Yes

- Positive diagonal gates give a matrix similar to a symmetric one.
- For $\beta \leq 1$, the delta factor is positive semidefinite.
- A scalar gate always leaves the transition symmetric.

Paper, Sec. 3, PDF p. 4; Theorem 8, App. A.3, pp. 22–23.

Two reflections make a rotation in one update

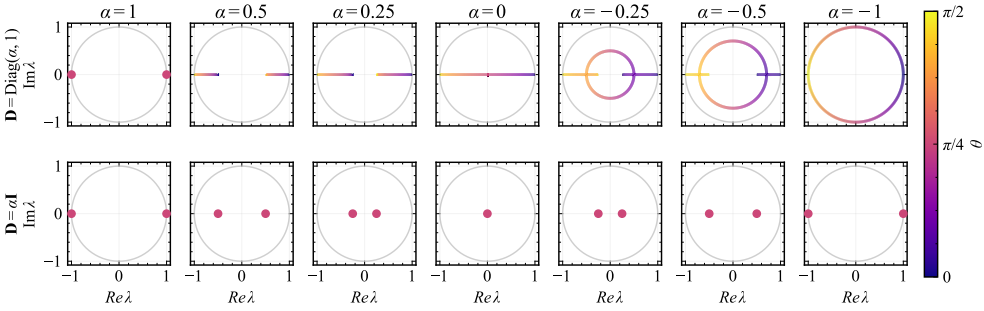


$$k = \frac{1}{\sqrt{2}}\begin{pmatrix}1\\1\end{pmatrix}, \quad (I - 2kk^\top)\begin{pmatrix}-1 & 0\\0 & 1\end{pmatrix} = \begin{pmatrix}0 & -1\\1 & 0\end{pmatrix} = R_{\pi/2}.$$

- The signed gate supplies a coordinate reflection.
- The delta factor supplies the second, freely oriented reflection.
- Real-valued computation; the name refers to a complex spectrum.

Paper, Fig. 1, PDF p. 2; Eq. (2), p. 4. Matrix expansion is explanatory.

Negative channel gates open a path to the unit circle



$$\Delta = (1-\alpha)^2\cos^2(2\theta) + 4\alpha. \qquad \Delta < 0 \implies |\lambda| = \sqrt{-\alpha}.$$

Top: $D = \mathrm{Diag}(\alpha, 1)$. Bottom: scalar $D = \alpha I$.

Paper, Fig. 2, PDF p. 2; Sec. 3, p. 4. Here $\beta = 2$ and $k = (\cos\theta, \sin\theta)^\top$.

The transition remains non-expansive and inexpensive

$$\|A_t\|_2 \le \underbrace{\left\|I - \beta_t k_t k_t^\top\right\|_2}_{\le 1} \underbrace{\|\mathrm{Diag}(\alpha_t)\|_2}_{\le 1} \le 1.$$

Same structural budget

$$A_t = D_t - \beta_t k_t (k_t^\top D_t) \quad \Rightarrow \quad \text{diagonal + rank one.}$$

State and recurrent work per token remain $O(nd_v)$ per head.

The bound controls homogeneous state propagation; repeated additive writes can still increase the state norm.

Paper, Secs. 2–4, PDF pp. 3–6; App. C.6, p. 41. Norm bound expanded for explanation.

CKDA spans the entire orthogonal DPR1 family

Theorem 1

Every real orthogonal $A = D + uv^\top$ with diagonal D can be written

$$A = (I - 2kk^\top)S, \qquad \|k\|_2 = 1, \quad S_{ii} \in \{-1, +1\}.$$

- A fully asymmetric rank-one correction adds no orthogonal transitions.
- Any CKDA transition has at most one non-real conjugate pair.
- General non-expansive DPR1: at most one such pair *on the unit circle*.

Paper, Theorem 1, PDF p. 5; Theorems 8–9, pp. 22–24. DPR1 = diagonal plus rank one.

One layer tracks all finite subgroups of $SO(3)$

Target group	Transition dimension	Construction
Cyclic $\mathbb{Z}_m$	2	Planar rotations
Dihedral D_m, including S_3	2	Rotations + reflections
A_4, S_4	3	Reoriented cube symmetries
A_5	4	Lift + many-to-one decoder

- **Realization**: one matrix for each group element.
- **Tracking**: multiple hidden matrices may decode to the same element.
- Constructions use prescribed initial states and exact algebraic values.

Paper, Theorem 3, PDF pp. 6–7; App. B–C, pp. 25–36. Dimension 4 uses a 4×4 matrix state.

A single layer still cannot track all of S_5

Theorem 4: the assumptions matter

Non-expansive transitions, at most one unit-circle complex pair per head, and finite reachability under *exact* affine updates.

- The obstruction allows any finite number of independent heads and an arbitrary joint decoder.
- Three CKDA layers suffice for every finite-group word problem.
- Regular languages and weighted finite automata (WFAs) over $\mathbb{Q}$ need $\beta > 2$ in this construction.

One-layer lower bound + three-layer construction leaves two layers open.

Paper, Theorems 4–5, PDF p. 7; precision in App. B.4, pp. 28–29; App. D.5, pp. 44–45.

Absorb cumulative signs into keys and queries

$$\mathrm{Diag}(\alpha_t) = S_t D_t, \qquad D_t = \mathrm{Diag}(|\alpha_t|), \quad P_t = S_1 \cdots S_t.$$
$$\widetilde{H}_t = P_t H_t, \quad \widetilde{k}_t = P_t k_t, \quad \widetilde{q}_t = P_t q_t.$$

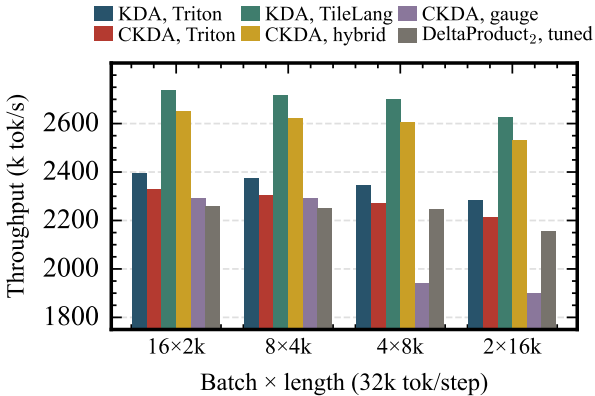
Equivalent positive-magnitude recurrence

$$\widetilde{H}_t = (I - \beta_t \widetilde{k}_t \widetilde{k}_t^\top) D_t \widetilde{H}_{t-1} + \beta_t \widetilde{k}_t v_t^\top, \quad o_t = \widetilde{H}_t^\top \widetilde{q}_t.$$

Reuse KDA kernels; fuse sign application with normalization and gradient writes.

Paper, Lemma 27 and implementation details, App. E, PDF pp. 45–46.

Fused CKDA retains 96–97% of KDA kernel throughput

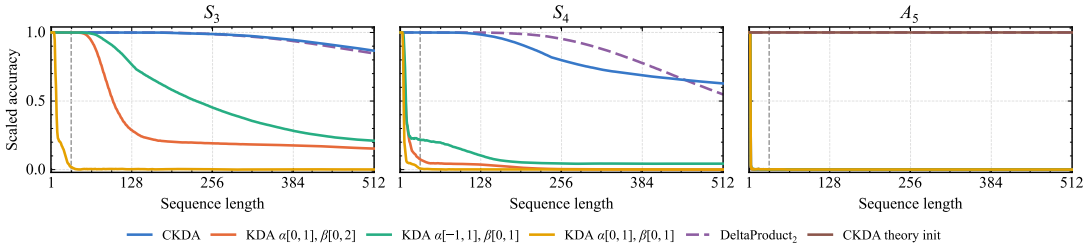


H100, BF16
16 heads, $d_k = d_v = 128$
32,768 tokens/step

- Forward + backward recurrence.
- Projections and optimizer excluded.
- Ungated DeltaProduct$_2$ remains competitive.

Paper, Fig. 5, PDF p. 7; App. E, p. 46. Gauge is an earlier implementation reference, not a matched fusion ablation.

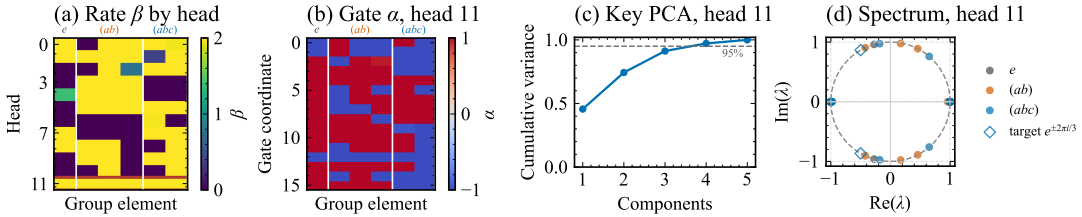
CKDA extrapolates better on S_3 and S_4



- Train to length 32; report best of three seeds.
- Full CKDA beats the other KDA ranges; S_4 still degrades.
- A_5 succeeds only in a separate theory-initialized setup.

Paper, Fig. 6 and Sec. 6, PDF p. 8; App. G.2, p. 47. Scaled accuracy maps chance to 0 and perfection to 1.

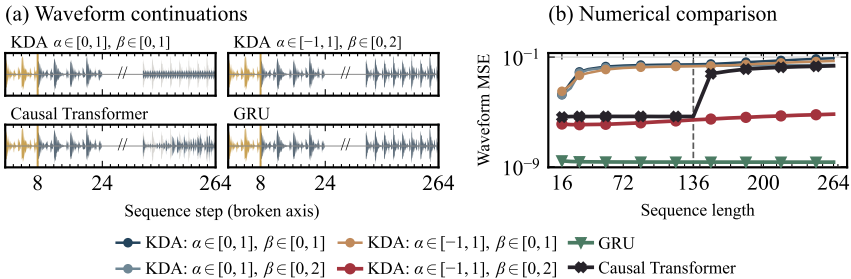
Successful state trackers learn the predicted geometry



- A successful head approaches $\beta = 2$ and sign-valued gates.
- Key PCA: the first three components explain about 95% of variance.
- Complex eigenvalues appear near the unit circle.

Paper, Fig. 7, PDF p. 8. The A_5 theory-initialized run also changes architecture and schedule, so initialization is not isolated.

Periodic continuation tests phase memory without feedback



- One fixed groove, 32 phases; 8-frame cue, then zero inputs.
- Train to length 136; at 264: 38.1 dB CKDA vs. 2.8 dB Transformer.
- CKDA wins among KDA range settings; the nonlinear GRU has lower MSE.

Paper, Fig. 8 and Sec. 6, PDF p. 9; App. G.3, pp. 47–48. Single-layer models; one training seed.

Small-model gains depend on initialization

340M / 15B Nemotron-CC tokens	Val. PPL ↓	Avg. accuracy ↑
Dense Attention	17.99	50.90%
Standard KDA	**17.45**	51.32%
CKDA, standard init	17.74	51.10%
CKDA, gate spread	17.49	**52.30%**
CKDA, spread, no key SiLU	17.49	51.87%

- Best CKDA gains 0.98 percentage points over KDA.
- Its validation perplexity is slightly worse.
- This average uses ten tasks; the 1.3B evaluation uses nine.

Paper, selected rows of Table 6, PDF p. 50; architecture details in Table 4, p. 49.

At 1.3B, matched CKDA and KDA controls perform similarly

Model	Wiki. PPL ↓	LAMBADA PPL ↓	Avg. acc. ↑
KDA, bounded gate	**15.73**	10.53	**54.09**
CKDA	15.78	**10.08**	54.06
KDA + attention 3:1	**15.04**	9.93	53.92
CKDA + attention 3:1	15.34	**9.90**	**54.37**

- Recurrent: -0.03 pp; hybrid: $+0.45$ pp in nine-task accuracy.
- The authors' KDA control also uses a bounded safe sigmoid gate.
- Their 3:1 full-attention hybrid differs from cited 1:1 sliding-window hybrids.

Paper, Table 2, PDF p. 9; comparison details, pp. 9–10. 100B FineWeb-Edu tokens; accuracy in percent.

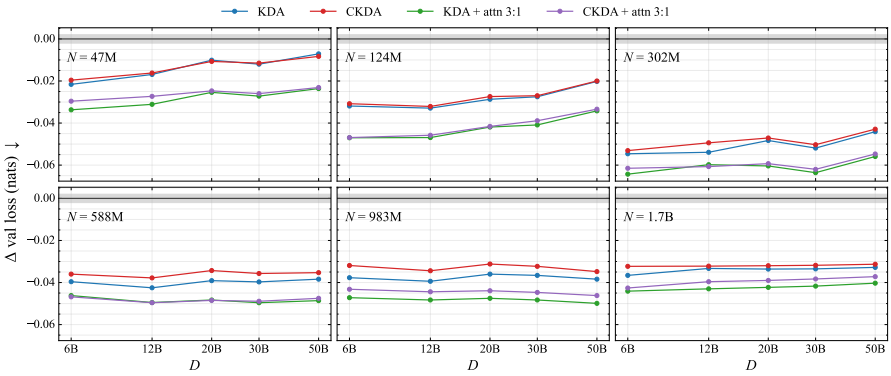
Retrieval results show tradeoffs across tasks

Recall (%) ↑	SQuAD	SWDE	FDA
KDA	38.17	**49.59**	30.40
CKDA	**39.68**	48.42	**31.49**
KDA + attention 3:1	**42.69**	**64.81**	**64.79**
CKDA + attention 3:1	37.77	63.91	58.89

- Hybrid accuracy gains do not extend to these three recall metrics.
- RULER is also mixed and highly sensitive to prompt wording.
- The paper attributes some low RULER scores to instruction-following failures.

Paper, Table 2, PDF p. 9; RULER discussion and Table 7, pp. 49–50.

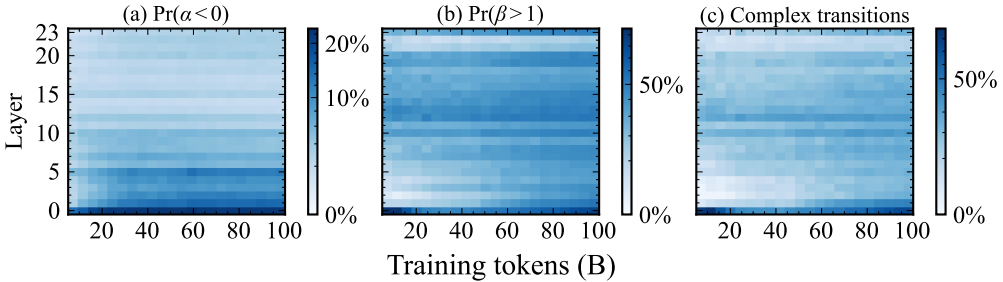
Scaling favors the recurrent family, with KDA close to CKDA



ΔL relative to the QK-Norm Transformer: more negative is better.
Across 47M–1.7B and 6B–50B tokens, KDA often has slightly lower loss.

Paper, Fig. 9, PDF p. 10; full loss grid, Table 12, p. 53. Nemotron-CC, context 4096.

Language models learn to use the newly allowed dynamics



- Negative gates emerge from positive standard initialization.
- $\beta > 1$ appears across layers; complex pairs concentrate early.
- Functional roles in language remain an open mechanistic question.

Paper, Fig. 10, PDF p. 10. 1.3B / 100B FineWeb-Edu tokens; colors use square-root scaling.

The contribution is more expressive dynamics at a small cost

1. **Mechanism:** signed channel gates + extended delta rate create rotations within a real DPR1 transition.
2. **Theory:** full orthogonal DPR1 coverage and stronger group tracking, with explicit precision and reachability assumptions.
3. **Evidence:** better symbolic and periodic extrapolation; general language modeling remains close to matched KDA.

Representability, learnability, and practical numerical robustness are distinct.

Paper, Sec. 7, PDF pp. 10–11; App. B.4, pp. 28–29.

Paper and discussion resources

Complex KDA: Understanding and Enhancing the Expressivity of Kimi Delta Attention

Julien Siems, Riccardo Grazzi, Korbinian Pöppel, Jaisidh Singh, Arber Zela, Timur Carstensen, Jenia Jitsev, Frank Hutter, Volkan Cevher, Antonio Orvieto, Aaron Klein.

- arXiv:2609.24797v1, submitted 21 September 2026.
- Author code: OpenEuroLLM/ComplexKDA
- Author model collection

Main discussion: approximately 25 minutes.
The following slides contain derivations, assumptions, and additional results.

Figures are original paper assets or an exact PDF crop. Tables select reported rows; explanatory derivations are labeled.

Backup: separate the sign core from magnitude scaling

$$A = \widetilde{A}\,\mathrm{Diag}(|\alpha|), \qquad \widetilde{A} = (I - \beta k k^\top)S.$$

Decompose $k = k_- + k_+$ into the negative/positive gate subspaces, with $\|k_-\| = \cos\theta$, $\|k_+\| = \sin\theta$.

$$\widetilde{A}\Big|_{\mathrm{span}(k_-,k_+)} = \begin{pmatrix} \beta\cos^2\theta - 1 & -\beta\sin\theta\cos\theta \\ \beta\sin\theta\cos\theta & 1 - \beta\sin^2\theta \end{pmatrix}.$$

- At $\beta = 2$, this block is a 2θ rotation; its complement has ± 1.
- Complex pair iff $\beta^2\cos^2(2\theta) < 4(\beta - 1)$.
- Nonuniform magnitudes need a separate spectral argument for full A.

Paper, Proposition 2, PDF p. 6; App. A.2–A.3, pp. 21–23. Both key components assumed nonzero.

Backup: the four-dimensional A_5 tracker uses a decoder

For a 3D rotation $R_{u,\theta}$, encode

$$a = \begin{pmatrix} \cos(\theta/2) \\ \sin(\theta/2)u \end{pmatrix}, \qquad g(R) = (I_4 - 2aa^\top)\,\mathrm{Diag}(-1,1,1,1).$$

$$Q = \begin{pmatrix} c & r^\top \\ s & C \end{pmatrix} \quad \Longrightarrow \quad f(Q) = cC - sr^\top - C[r]_\times.$$

The required invariant

$$f(g(R)Q) = R\,f(Q), \qquad f(I_4) = I_3.$$

At most $2|A_5|^2 = 7200$ hidden matrices; several can encode the same group element.

Paper, Proposition 18 and Corollary 19, App. C.4.2, PDF pp. 34–36. $[r]_\times x = r \times x$.

Backup: three layers implement clock, buffer, accumulator

1. **Clock:** one planar rotation tracks position modulo $2m$.
2. **Buffer:** retain the last $2m$ inputs and select block factors.
3. **Accumulator:** apply one factor per step; complete the delayed product in the readout.

Target	Sufficient factor budget	Rate condition
Permutation on n coordinates	$n-1$	$\beta \in [0,2]$
Explicit DFA transition	$2n-2$	$\beta \in [0,3]$
Rational matrix, with scratch	$7n^2+3n$	$\beta \geq 0$, unbounded

Paper, App. D, PDF pp. 41–45; $n \geq 2$ for the first two budgets. Exact algebraic arithmetic; WFA storage grows polynomially in length.

Backup: practical gates differ from exact theoretical endpoints

$$r = \tanh(a/2), \quad \alpha = s[\epsilon + (1-\epsilon)|r|], \quad \epsilon = e^{-5}, \quad \beta = 2\sigma(b).$$

- $s = +1$ for $r \geq 0$ and -1 otherwise; $\log|\alpha|$ stays in $[-5, 0]$.
- Gate is discontinuous at zero; backward detaches s and uses the zero subgradient of $|r|$ at zero.
- Standard init: positive gates. Spread init: about half negative.
- Finite sigmoid inputs approach but do not attain $\beta = 2$, $|\alpha| = 1$.

Removing key SiLU changes an optimization bias, not reachable key directions.

Paper, App. E–G.1, PDF pp. 45–47.

Backup: RULER gains depend on task and architecture

Accuracy at 4K (%)	S-NIAH-1	S-NIAH-2	S-NIAH-3	MK-NIAH-1
KDA	94.0	**97.6**	9.8	26.0
CKDA	**100.0**	88.6	**51.0**	**37.0**
KDA + attn 3:1	**100.0**	**99.2**	**95.8**	63.8
CKDA + attn 3:1	81.2	83.2	86.8	**94.4**

- 500 samples per cell; 4K training context.
- Prompt variations can strongly affect instruction following.
- Table 7 says hybrid evaluation stops at 4K, yet includes 8K cells. Only 4K hybrid cells are compared here.

Paper, Table 7, PDF p. 50; prompt-sensitivity discussion, p. 49. Caption/data conflict is in v1.

Backup: the largest scaling cell has mixed rankings

1.7B / 50B Nemotron-CC tokens	Loss (nats) ↓	Avg. acc. (%) ↑
Transformer++ (QK-Norm)	2.1374	55.10
KDA	2.1046	56.87
CKDA	2.1061	**57.21**
KDA + attn 3:1	**2.0971**	56.47
CKDA + attn 3:1	2.1002	55.82

$$L(N, D) = E + (AN^{-\alpha_s} + BD^{-\beta_s})^k.$$

Fit holds out this cell. Free-E CKDA error: $+0.0007$ nats.
Shared training trunks and fixed hyperparameters limit uncertainty claims.

Paper, Tables 11–13, PDF pp. 52–54; fitting methodology, p. 51. Some lower-scale accuracy tasks are near chance.

Backup: why the S_5 impossibility needs finite reachability

1. Finite exact reachability + non-expansion permits an orthogonal reduction that removes additive terms.
2. The finite hidden matrix group maps onto S_5 through the decoder.
3. Lift a five-cycle and its transposition conjugate to per-head planar rotations.
4. Their matrix commutator satisfies $C^{10} = I$, while its decoded three-cycle does not.

Arithmetic boundary

A finite floating-point format does not imply finite reachability for the exact affine recurrence. Rounding changes the update map.

Paper, App. B.4, PDF pp. 28–29; App. C.5–C.6, pp. 36–41.