CyFA: Linear Sequence Modeling with Relative-Time-Partitioned Memory

Research paper discussion (25–30 minutes + backup)

Paper authors: Yixiao Chen, Shuojin Yang, Shi-Min Hu

Reading presentation prepared by Codex

arXiv:2609.36259v1 — September 28, 2026

Reading date: October 2, 2026

Organize a fixed memory by model age; compute it with two scalar recurrences.

More history competes for the same recurrent state

$$S_t = S_{t-1} + k_t v_t^\top, \qquad o_t = S_t^\top q_t = \sum_{i \le t} v_i (k_i^\top q_t).$$

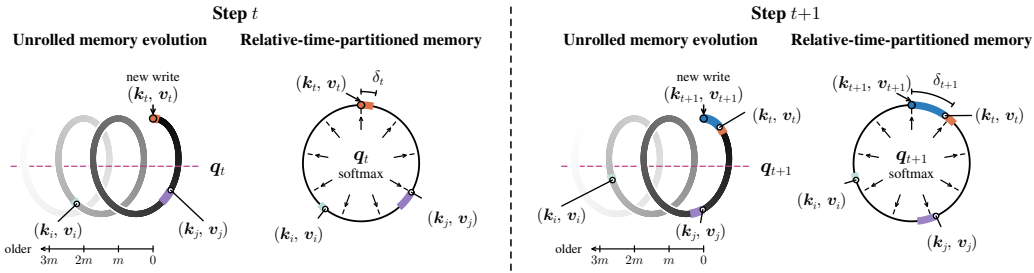
- A fixed matrix compresses every earlier association.
- Forget gates reduce persistence; Delta Rule updates correct the association addressed by a new key.
- Multiple useful associations can still interfere at retrieval.

CyFA's question

Can the same state budget keep associations more distinguishable by organizing their temporal addresses?

Paper, Secs. 1–2, PDF pp. 1–3; Eq. (1).

A learned clock moves old writes around fixed slots



- New KV enters age zero; old KV moves by $\delta_t \in (0, 1)$.
- The unrolled helix is a visualization: every cycle reuses the same slots.
- Scalar decay controls how much old content survives.

Paper, Fig. 1, PDF p. 2. Original figure; R and write strength are omitted in the illustration.

Transport, forgetting, and writing have separate controls

$$\lambda_t = \lambda_{t-1} + \delta_t, \qquad K_t \in \mathbb{R}^{m \times d_k}, \quad V_t \in \mathbb{R}^{m \times d_v}$$

$$K_t = \alpha_t P(\delta_t) K_{t-1} + \beta_t e_0 k_t^\top,$$

$$V_t = \alpha_t P(\delta_t) V_{t-1} + \beta_t e_0 v_t^\top.$$

- $P(\delta_t)$: joint cyclic movement of aligned K/V states.
- α_t: scalar retention; β_t: current write strength.
- e_0: age-zero slot; $\lambda_t - \lambda_i$: model age of write i.

$$o_t = (RV_t)^\top \operatorname{softmax}(RK_t q_t), \qquad R \in \mathbb{R}^{m \times m}.$$

Paper, Eqs. (14)–(18), PDF p. 5. R is learned, initialized to I; conventional dot-product scaling omitted as in the paper.

Fractional shifts interpolate between integer slots

$$P(\tau) = \Phi \mathcal{U}(\tau) \Phi^{\top}, \qquad \mathcal{U}(\tau) = 1 \oplus \bigoplus_{j=1}^{(m-1)/2} \mathrm{Rot}\left(\frac{2\pi j \tau}{m}\right).$$

- Φ: orthogonal real Fourier basis; m is odd.
- $P(1)$ is an exact cyclic shift; $P(m) = I$.
- $P(a)P(b) = P(a+b)$: clock increments add up.

$$e_r^{\top} P(\tau) e_s = \frac{\sin\big(\pi(\tau + s - r)\big)}{m \sin\big(\pi(\tau + s - r)/m\big)}.$$

A geometric interpolation, not a probability distribution

Coefficients can be negative. Transport preserves norms, but repeated cycles still overlap.

Paper, Eqs. (12)–(13), PDF p. 4; App. A.1–A.2, Eq. (39), PDF pp. 15–16.

Changing coordinates cancels the recurrent rotation

$$b = \Phi^\top e_0, \qquad \overline{K}_t = \mathcal{U}(-\lambda_t)\Phi^\top K_t.$$

The cancellation

$$\mathcal{U}(-\lambda_t)\mathcal{U}(\delta_t) = \mathcal{U}(-\lambda_{t-1}).$$

$$a_t = \beta_t \mathcal{U}(-\lambda_t) b, \qquad \begin{aligned} \overline{K}_t &= \alpha_t \overline{K}_{t-1} + a_t k_t^\top, \\ \overline{V}_t &= \alpha_t \overline{V}_{t-1} + a_t v_t^\top. \end{aligned}$$

- Existing states now undergo only scalar decay.
- Clock dependence moves into write/read vectors.
- Exact reconstruction: $K_t = \Phi \mathcal{U}(\lambda_t)\overline{K}_t$, and likewise for V_t.

Paper, Eqs. (19)–(25), PDF pp. 5–6. a_t abbreviates the paper's temporal write vector.

Two scalar attention passes surround a slot softmax

$$\{o'_t\} = \mathrm{ScalarGatedLA}(q, k, a, \alpha), \qquad o'_t = \overline{K}_t q_t$$

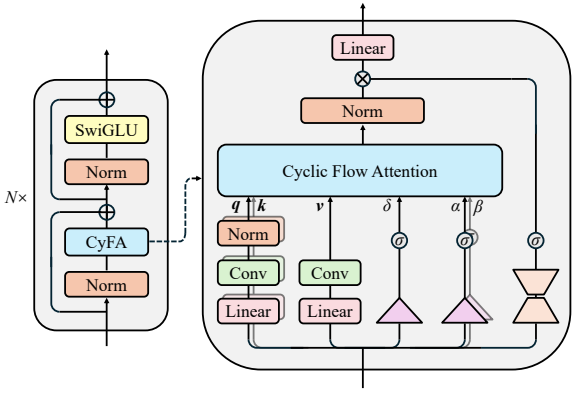
$$F_t = R\Phi\mathcal{U}(\lambda_t), \qquad o''_t = F_t^\top \,\mathrm{softmax}(F_t o'_t)$$

$$\{o_t\} = \mathrm{ScalarGatedLA}(o'', a, v, \alpha), \qquad o_t = \overline{V}_t^\top o''_t$$

- Pass 1: content matches accumulate evidence in temporal addresses.
- Bridge: token-wise transformations on m-dimensional vectors.
- Pass 2: the address query retrieves associated values.
- Training uses clock prefix sums and chunked recurrences; decoding carries fixed states and clocks.

Paper, Eq. (27), PDF p. 6; App. B, PDF pp. 20–21. Softmax is over memory slots, not all past tokens.

The mixer combines CyFA with established network components



Paper, Fig. 2 and Sec. 3.4, PDF p. 7; App. A.7, PDF p. 20.

- ShortConv on Q/K/V.
- RMSNorm on Q/K and head outputs.
- Low-rank sigmoid output gate.
- Pre-norm backbone with SwiGLU.
- $m = 127$ active slots; $m^* = 128$ allocated rows.

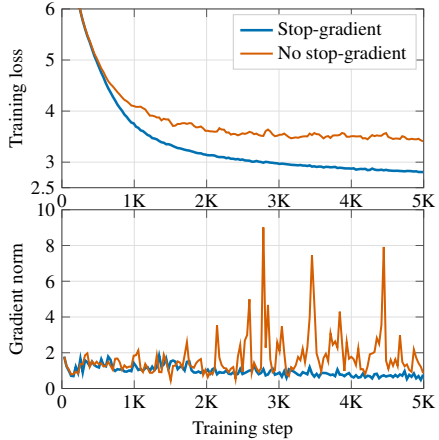
The padding row is masked and stays zero; the cycle has 127 slots.

A stop-gradient protects the backbone from clock gradients

$$\delta_t = \sigma(W_\delta\,\mathrm{sg}(x_t) + b_\delta)$$

$$\frac{\partial \mathcal{L}}{\partial \delta_s} = \sum_{t \geq s} \frac{\partial \mathcal{L}}{\partial \lambda_t}$$

- Each increment receives future clock-path gradients.
- Stop-gradient blocks this branch's gradient into x_t.
- W_δ, b_δ still learn; forward values stay input dependent.



Paper, Eq. (28), Fig. 3, PDF p. 7; Eq. (95), App. D.1, PDF pp. 23–24. Curves: first 5,000 steps.

The controlled comparisons match main recurrent state size

Scale	Training tokens	CyFA parameters	State entries/layer
400M	15B	389M	262,144
800M	30B	800M	393,216
1.4B	100B	1,394M	524,288

$$\underbrace{4 \cdot 128 \cdot (256 + 256)}_{\text{CyFA, 400M}} = \underbrace{4 \cdot 256 \cdot 256}_{\text{GDN/KDA, 400M}} .$$

- 24 layers, SlimPajama, Mistral tokenizer, initial 2K context.
- State count excludes ShortConv cache; total memory and parameter counts are not exactly equal.
- Public 1.4B checkpoints are a separate, unmatched reference group.

Paper, Sec. 4 and App. C, PDF pp. 8, 22–23; Tables 6–7.

Evaluation separates recall, general quality, and efficiency

General quality	Wiki/LAMBADA perplexity ↓; six zero-shot commonsense tasks, average score ↑.
Recall at 2K	FDA, SWDE, SQD, NQ, TQA, DROP; protocol from `prefix-linear-attention`; inputs truncated to 2K.
Long inputs	LongBench at 8K; RULER at 1K–8K. Main results follow 1,024 continuation steps at 8K.
Efficiency	End-to-end training tokens/s and core-operator latency are measured separately.

Metric boundary: LongBench averages task scores; it is not one shared accuracy metric. No multi-seed error bars are reported.

Paper, Secs. 4.1–4.3, PDF pp. 8–10; App. C.2, PDF pp. 22–23.

Recall gains over KDA persist across three model scales

Scale	KDA	CyFA	Gain (percentage points)
400M	26.31	33.25	+6.94
800M	31.18	37.13	+5.95
1.4B	37.73	43.47	+5.74

What this result supports

Better retrieval from a matched main-state budget across the reported training scales.

- The average covers six real-world recall-intensive tasks at 2K.
- CyFA leads the recurrent models on every recall task at each scale.
- Full Transformers still lead the average at 400M and 800M.

Paper, Table 2, PDF p. 8. Gains are computed from reported averages.

Field-value retrieval shows the largest 400M gains

Model	FDA	SWDE	SQD	Recall avg.
Transformer	53.04	38.89	30.70	33.98
Mamba-3	24.89	27.84	29.53	26.46
KDA	26.07	25.02	29.49	26.31
CyFA	42.60	35.61	34.63	33.25

- FDA: +16.53 points over the best other recurrent model, KDA.
- SWDE: +7.77 points over the best other recurrent model, Mamba-3.
- Many coexisting field-value pairs directly stress compressed associative memory.

The full Transformer remains stronger on FDA and SWDE at this scale.

Paper, Table 2 and Sec. 4.1, PDF pp. 8–9. All scores ↑; avg. uses all six recall tasks.

General quality is competitive, with specific trade-offs

Scale	Model	Wiki PPL ↓	Lamb. PPL ↓	Common avg. ↑
400M	KDA	25.76	51.17	41.90
	CyFA	25.85	48.17	42.47
800M	KDA	19.83	22.20	46.81
	CyFA	19.51	18.93	48.16
1.4B	KDA	15.97	11.88	52.89
	CyFA	15.47	12.77	53.19

- Commonsense averages improve at all three scales.
- Wiki PPL is slightly worse at 400M; LAMBADA PPL is worse at 1.4B.
- Small differences have no reported uncertainty estimates.

Paper, Table 2, PDF p. 8. Selected rows transcribed; no uniform per-metric dominance is claimed.

LongBench improves after 8K continuation

Scale	Model	Average	LCC	RBP
800M	KDA	15.47	41.47	39.06
	CyFA	17.19	47.84	43.59
1.4B	KDA	20.02	46.34	42.56
	CyFA	21.64	53.60	46.40

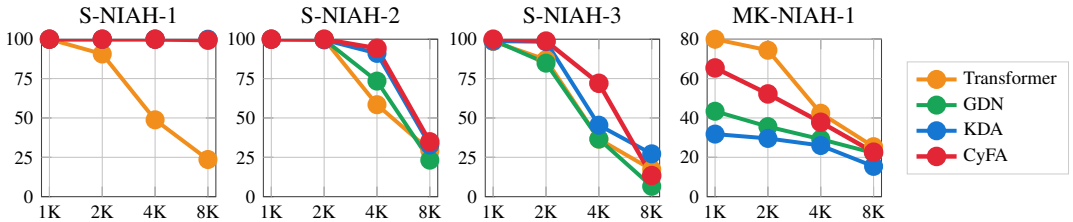
- Inputs truncated to 8K; 1,024 extra training steps at 8K.
- Both code tasks improve. At 1.4B, LCC gains 7.26 points.
- At 1.4B, CyFA trails KDA on HQA, TRE, and SAM.

A counting inconsistency in the paper

Text says 15 tasks; the table has 14 task columns plus Avg. CyFA leads 10/14 at 800M and 11/14 at 1.4B.

Paper, Table 3 and Sec. 4.2, PDF p. 9. Average recomputed over the 14 displayed tasks.

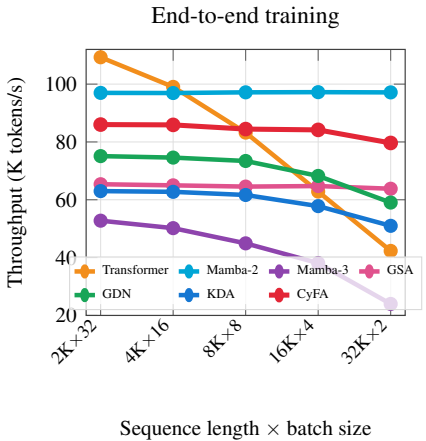
Multi-key retrieval improves, but long inputs still degrade



- 800M models, after 8K continuation; x-axis: 1K–8K tokens.
- MK-NIAH-1 requires multiple associations to coexist: CyFA clearly leads GDN/KDA through 4K.
- At 8K, difficult retrieval remains weak; superiority is not uniform across all panels and lengths.

Paper, Fig. 4, PDF p. 9. Original figure; the multi-key panel has a different y-axis scale. No exact points inferred from the plot.

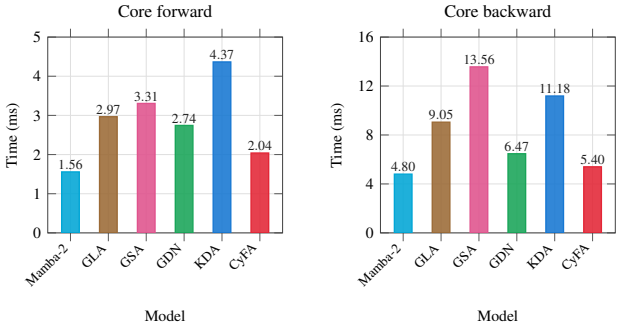
End-to-end training reaches 1.37× KDA at 8K



- 400M models on one RTX Pro 6000.
- Same tokens per batch: longer sequences use smaller batches.
- CyFA stays around 80–86K tokens/s from 2K to 32K.
- Faster than the Transformer from 8K onward.
- Mamba-2 is faster than CyFA among recurrent models shown.

Paper, Fig. 5 and Sec. 4.3, PDF pp. 9–10. Training throughput, not autoregressive serving throughput.

Core operator latency is about half of KDA's



	KDA	CyFA	CyFA / KDA
Forward	4.37 ms	2.04 ms	46.7%
Backward	11.18 ms	5.40 ms	48.3%

Paper, Fig. 6, PDF p. 10. One RTX Pro 6000; 400M training shapes, sequence length 2K, batch 32. These ratios describe the core operators only.

Forgetting, adaptive clocks, and capacity all matter

400M variant	Lamb. PPL ↓	Recall avg. ↑	Δ recall
Default, $m = 127$	48.17	33.25	–
No forgetting ($\alpha = 1$)	115.89	28.05	−5.20
Fixed increment per head	51.98	30.32	−2.93
Fixed readout ($R = I$)	58.75	31.79	−1.46
No QK normalization	51.37	30.12	−3.13
$m = 63$	51.34	26.68	−6.57
$m = 255$	46.79	34.83	+1.58

Each row is a separate change to the default. More slots also increase the memory/computation budget.

Additional trade-off: adding SiLU improves PPL but reduces recall to 32.26.

Paper, Table 4, PDF p. 10. Full table and ShortConv ablation are included in the Chinese notes.

Clock interventions support contextual alignment

- 400M: 56/96 heads are near unit increments; 40/96 are input dependent.
- Interventions affect the 40 input-dependent heads; 16 held-out sequences.

Clock setting	NLL ↓	ΔNLL
Original	2.6105	–
Per-head mean	2.6555	+0.0450
Shifted by 257 tokens	2.7344	+0.1238
Set to one	2.7486	+0.1381

Interpretation

Shifting preserves the increment distribution but breaks alignment with inputs. Its larger penalty supports use of context-dependent timing.

Paper, Sec. 4.5, PDF p. 10; App. D.2 and Table 8, PDF p. 24. A single-checkpoint intervention, not a universal semantic-importance measure.

The contribution is better organization within finite capacity

1. **Representation:** a learned clock controls temporal spacing while K/V remain aligned.
2. **Computation:** an exact coordinate change yields two scalar-decay recurrences.
3. **Evidence:** recall improves at matched main-state sizes with competitive training throughput.

Boundaries to retain

Cycles still alias; difficult 8K retrieval degrades. Some metrics regress. No multi-seed uncertainty or serving-latency benchmark is reported.

Discussion: How should clock resolution and slot capacity be allocated across heads under a total memory and compute budget?

Synthesis of Secs. 3–4 and Apps. A, C–E. The discussion question is this reading's analysis, not a paper result.

Source, reading scope, and verification

Yixiao Chen, Shuojin Yang, and Shi-Min Hu.
CyFA: Linear Sequence Modeling with Relative-Time-Partitioned Memory.
arXiv:2609.36259v1, September 28, 2026.

- Versioned arXiv page Paper PDF
- Author-linked code: github.com/Chyxx/CyclicFlowAttention.
- Reading covers all main sections and Appendices A–E; figures are copied from the matching author TeX source.
- Independent small FP64 check: state reconstruction error $< 4 \times 10^{-15}$; output error $< 8 \times 10^{-16}$.
- This is an equation check, not a reproduction of training, gradients, or the authors' kernels.

All reported benchmark numbers come from this paper. LongBench task-count inconsistency is identified explicitly. Backup slides follow.

Backup: temporal overlap factors out from content similarity

Let $p_i(t) = P(\lambda_t - \lambda_i)e_0$ and $D_m(x) = \sin(\pi x)/[m \sin(\pi x/m)]$.

$$p_i(t)^\top p_j(t) = D_m(\lambda_i - \lambda_j)$$

$$\langle p_i(t)k_i^\top, p_j(t)k_j^\top \rangle_F = (k_i^\top k_j) D_m(\lambda_i - \lambda_j).$$

- Nonzero integer separations within one cycle are orthogonal.
- Fractional separations generally overlap; full-cycle separations alias.
- Later shared transport preserves pairwise profile geometry.
- This identity excludes scalar write/forget amplitudes and is not a recall guarantee.

Paper, App. A.3, Proposition 4 and Corollary 1; Eqs. (54), (60), PDF pp. 17–18.

Backup: scalar decay preserves a matrix-multiply chunk form

For a chunk of C tokens, define $\gamma_r = \prod_{j=1}^{r} \alpha_j$ and

$$\Gamma_{rs} = \begin{cases} \gamma_r/\gamma_s, & r \geq s, \\ 0, & r < s. \end{cases}$$

For $S_t = \alpha_t S_{t-1} + k_t v_t^\top$,

$$S_{\text{out}} = \gamma_C S_{\text{in}} + K^\top \operatorname{Diag}(\gamma_C/\gamma_r) V,$$

$$O = \operatorname{Diag}(\gamma) Q S_{\text{in}} + (QK^\top \odot \Gamma) V.$$

- First term: earlier chunks. Second term: causal interactions within the chunk.
- Key pass uses $(Q, K, V) = (q, k, a)$; value pass uses (o'', a, v).
- The two passes share decay factors; use chunk-local sums of $\log \alpha$.

Paper, App. B, Eqs. (83)–(88), PDF p. 21. Symbols here are local to one chunk.

Backup: constant clocks recover BLA's temporal profile

Set $\delta_t = 1/\rho$. Then model age is $(t-i)/\rho$ and one cycle spans $T = \rho m$ tokens.

$$[P((t-i)/\rho)e_0]_r = \frac{1}{m}\left[1 + 2\sum_{h=1}^{(m-1)/2} \cos\left(\frac{2\pi h(t-i-\rho r)}{T}\right)\right].$$

- This is the fixed Fourier–Dirichlet profile used by BLA.
- CyFA separately learns timing, scalar forgetting, and readout R.
- Shared profile does not make the full recurrences identical.
- The authors implement BLA through its GSA-equivalent recurrence, with the same surrounding mixer as CyFA.

Paper, App. A.6, Eqs. (79)–(80), PDF p. 20; App. C.1, PDF p. 22.

Backup: state matching is narrower than total-memory matching

400M family	Main state entries/layer	Count
CyFA/BLA	$Hm^*(d_k + d_v)$	262,144
SWA/GSA/Raven	$Hm(d_k + d_v)$	262,144
GDN/KDA	$Hd_k d_v$	262,144
Mamba-2/3	$Hd_{\text{head}} d_{\text{state}}$	262,144

- Slot models: $H = 4$, $d_k = d_v = 256$, allocated rows 128.
- Mamba: $H = 32$, $d_{\text{head}} = 64$, $d_{\text{state}} = 128$.
- ShortConv cache is excluded. Learned R is a parameter, not the main recurrent state.
- Public 1.4B checkpoints retain released architectures and tokenizers.

Paper, App. C.1, Eqs. (90)–(94) and Table 6, PDF pp. 22–23.

Backup: paired backward probes localize the clock gradient

Model / derivative setting	Global norm	Excluding clock projection
KDA reference	2.309	2.309
CyFA with stop-gradient	5.653	2.447
CyFA without stop-gradient	12.006	8.682

- Four 2K sequences from the 400M checkpoint; same forward values and loss for the paired CyFA probes.
- Clock projection accounts for 78.3% of the squared global norm with stop-gradient.
- Restoring the clock-to-input derivative spreads larger gradients into the backbone.
- This is a local gradient probe, not a general stability theorem.

Paper, App. D.1, PDF pp. 23–24. Table transcribes reported norms; KDA has no shared cumulative-clock projection.

Backup: continuation improves averages, not every task

Scale	Model	Before 8K continuation	After 8K continuation
800M	Transformer	2.57	11.86
	GDN	14.13	14.57
	KDA	14.80	15.47
	CyFA	15.75	17.19
1.4B	KDA	19.05	20.02
	CyFA	20.47	21.64

- Both evaluations truncate LongBench inputs to 8K.
- Continuation: 1,024 steps, constant 3×10^{-5} learning rate, unchanged tokens per optimizer step.
- Example regression: 1.4B CyFA GvR, $12.58 \rightarrow 5.98$.

Paper, Tables 3 and 9, PDF pp. 9 and 25; App. C.2, PDF p. 23. Scores are averages over the 14 displayed tasks.