Fractional State Space Transition for Long Sequence Modeling

A reading seminar on FRAC

Paper authors: Ivan Kobyzev, Abbas Ghaddar, Ali Nasiri-Sarvi, Lifeng Shang, Yufei Cui

Huawei Noah's Ark Lab, Montreal Research Center

arXiv:2609.36314v1 | 28 September 2026
Reading prepared: 7 October 2026

A finite bank of exponential modes provides a practical long-memory prior.

A bounded state still needs a memory law

Research question

How can a recurrent model retain useful history across widely separated timescales?

$$\dot{h}(t) = -\lambda h(t) + u(t) \quad \Longrightarrow \quad h(t) = e^{-\lambda t} h(0) \quad \text{when } u = 0.$$

- A single exponential has one characteristic timescale, $1/\lambda$.
- FRAC organizes multiple decay modes using fractional dynamics.
- The goal is broader temporal coverage with a fixed-size decoding state.

Paper, Secs. 1–2.2, PDF pp. 1–2. Related-work distinctions: Sec. 5, p. 9.

Fractional dynamics induce two different tail laws

$$ {}^{C}D_t^{\alpha} h(t) = -\lambda h(t) + u(t), \qquad 0 < \alpha < 1, \quad \lambda > 0 $$

$$ h(t) = \underbrace{E_{\alpha}(-\lambda t^{\alpha}) h(0)}_{\text{initial-state relaxation}} + \underbrace{\int_0^t g_{\alpha,\lambda}(t-\xi) u(\xi)\, d\xi}_{\text{response to past inputs}} $$

Homogeneous kernel

$$ E_{\alpha}(-\lambda t^{\alpha}) \sim \frac{t^{-\alpha}}{\lambda \Gamma(1-\alpha)} $$

Impulse-response kernel

$$ g_{\alpha,\lambda}(t) \sim \frac{\alpha\, t^{-\alpha-1}}{\lambda^2 \Gamma(1-\alpha)} $$

These are large-time asymptotics. Direct fractional solvers retain history.

Paper, Sec. 2.3, Eqs. (4)–(8), PDF p. 3. *E* denotes a Mittag–Leffler function.

A geometric mode bank approximates fractional kernels

$$E_\alpha(-\lambda t^\alpha) = \int_0^\infty e^{-t/\tau} R_{\alpha,\lambda}(\tau)\, d\tau \approx \sum_{m=1}^{M} c_m e^{-t/\tau_m}$$

$$\tau_m = \tau_0 q^{m-1}, \qquad q > 1, \qquad c_m > 0.$$

- Positive diffusive density: fast and slow exponentials share the memory load.
- Log-space quadrature gives $c_m \propto \tau_m^{-\alpha}$ at large timescales.
- On a finite horizon $[0, T]$, Theorem 2 guarantees a suitable finite bank.

Different error norms are essential

Uniform error for the homogeneous kernel; L^1 error for the impulse kernel, which is singular at $t = 0$.

Paper, Theorems 1–2, Eqs. (9)–(12), PDF p. 4; Appendix B, pp. 17–19.

The mode bank has an exact local ZOH update

$$\dot{s}_m = -\frac{s_m}{\tau_m} + \frac{u}{\lambda \tau_m}, \qquad \tilde{h}_M = \sum_m c_m s_m$$

Hold the input and controls fixed during an interval of length Δ_t:

$$s_{t,m} = \rho_{t,m} s_{t-1,m} + \beta_{t,m} u_t, \quad \rho_{t,m} = e^{-\Delta_t/\tau_m}, \quad \beta_{t,m} = \frac{1-\rho_{t,m}}{\lambda}.$$

Token-dependent timescale control

$$\tilde{\tau}_{t,m} = \frac{\tau_m}{\lambda_t^{1/\alpha_t}}, \qquad \tilde{\rho}_{t,m} = \exp\left(-\frac{\Delta_t \lambda_t^{1/\alpha_t}}{\tau_m}\right).$$

The product $\Delta_t \lambda_t^{1/\alpha_t}$ advances an internal clock.

Paper, Proposition 3 and Eqs. (13)–(20), PDF p. 5; Appendix A, Remark 4, p. 17.

All experiments modify the theoretical write amplitude

Exact ZOH derivation

$$\tilde{\beta}_{t,m} = \frac{1 - \tilde{\rho}_{t,m}}{\lambda_t}$$

Algorithm 1: experimental model

$$\tilde{\beta}_{t,m} = 1 - \tilde{\rho}_{t,m}$$

- Authors remove the explicit $1/\lambda_t$ write-amplitude factor for training stability.
- Constant λ: related by constant state/readout rescaling.
- Token-varying λ_t: this is a substantive modeling choice.

Interpretation

The selective layer generalizes a fixed-system fractional construction; the theory supplies its memory prior.

Paper, Eq. (20), PDF p. 5; Algorithm 1 and Remarks 7–8, Appendix D, p. 21.

Read and write routing share a timescale prior

$$c_{t,m} = \mathrm{softmax}_m[-\alpha_t \log \tau_m + g_{\mathrm{read}}(u_t)_m],$$
$$b_{t,m} = \mathrm{softmax}_m[-\alpha_t \log \tau_m + g_{\mathrm{write}}(u_t)_m].$$

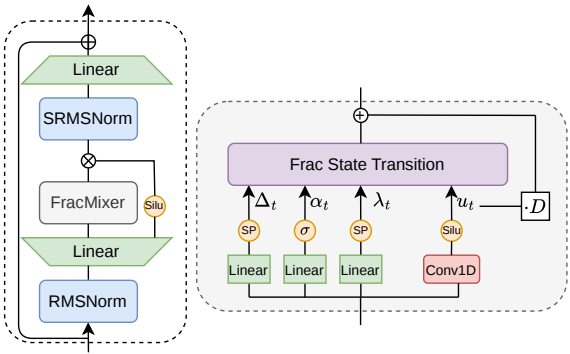
- The prior favors $\tau_m^{-\alpha_t}$; smaller α_t flattens this distribution.
- Independent content projections select where to write and read.
- The read prior has an asymptotic basis; matching the write prior is empirical.

Actual recurrent rule used in the experiments

$$s_{t,m} = \tilde{\rho}_{t,m} s_{t-1,m} + (1 - \tilde{\rho}_{t,m}) b_{t,m} u_t, \qquad h_t = \sum_m c_{t,m} s_{t,m}.$$

Paper, Sec. 3.2, Eq. (23), PDF p. 6; Algorithm 1, p. 21.

The mixer separates clock control from content



- Pre-convolution branch: Δ_t, α_t, λ_t.
- Content branch: causal depthwise convolution and SiLU.
- Direct feedthrough: $h_t + Du_t$.
- Gated normalization, output projection, residual connection.

$M = 16$, geometric $\tau \in [1, 2^{17}]$; $\Delta \in [10^{-4}, 1]$, $\lambda \in [0.25, 4]$.

Original Fig. 1, Sec. 3.3, PDF p. 6; language-model configuration, Appendix E.3, p. 23.

Affine recurrence supports scan and bounded-state decoding

$$\mathbf{s}_t = \Lambda_t \mathbf{s}_{t-1} + \mathbf{v}_t, \qquad \Lambda_t = \mathrm{Diag}(\tilde{\rho}_{t,1}, \ldots, \tilde{\rho}_{t,M}).$$

- Associative affine composition enables chunked parallel training and prefill.
- Decode uses the same one-step rule and caches $S_t \in \mathbb{R}^{H \times M \times d_h}$.
- Direct recurrence work: $O(LHMd_h)$; state per layer: $O(HMd_h)$.

Principal state per layer	Scalars	Relative to FRAC
FRAC	65,536	1×
Mamba2 / Mamba3	524,288	8×
GDN	294,912	4.5×

Paper, Eq. (24), p. 6; Appendix E.3, p. 23. Counts exclude convolution buffers and auxiliary caches.

The evaluation asks three different questions

Does the prior transfer across sequence lengths?

Heavy-tail probe: train at 512, test up to 128K.
Language-model NIAH: train at 4K, test up to 64K.

Does the model preserve useful general capabilities?

MADLab, LongBench, LM Harness, and real-world retrieval truncated to 2K.

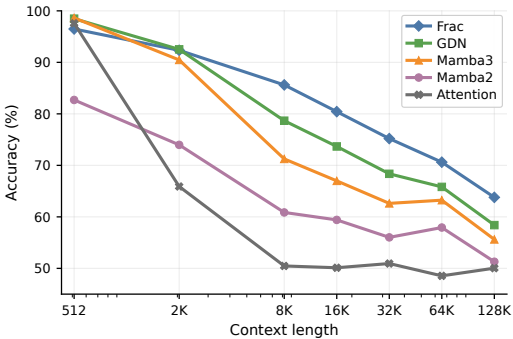
What are the costs and design dependencies?

Prefill / decode / training throughput; 390M model ablations; finite-bank fitting.

DNA uses a separate protocol: train and test at each chosen context length.

Paper, Sec. 4, pp. 6–9; Appendices E–G, pp. 21–27.

FRAC extrapolates the heavy-tail task more effectively



One layer, ∼200K parameters; train only at 512.

$$\text{label} = \text{sign} \sum_k \frac{v_k}{(L-k)^{0.1}}$$

128K accuracy, mean ± SD:

- FRAC: 63.8 ± 1.9
- GDN: 58.4 ± 3.8
- Attention: 50.0 ± 3.7

The 128K gain over GDN is 5.4 percentage points across 10 runs.

Original Fig. 2, p. 7; Appendix E.1, Tables 5–6, p. 22. Accuracy in percent.

MADLab shows preserved core abilities, with modest gains

Model	Comp.	F-ICR	Mem.	Copy	Avg.
Mamba2	88.3	7.2	99.9	99.9	73.8
GDN	88.9	9.7	98.8	99.9	74.3
Mamba3-SISO	89.1	10.1	100.0	100.0	74.8
FRAC	**90.4**	**11.5**	99.9	99.9	**75.4**

- $\sim$500K parameters; two sequence mixers and two SwiGLU layers.
- Compression and fuzzy recall account for the average gain.
- Fuzzy recall remains difficult: even the best accuracy is only 11.5%.
- Five-seed means after task-specific hyperparameter selection.

Paper, Table 1, PDF p. 7; Appendix E.2, pp. 22–23. Standard deviations omitted here for readability.

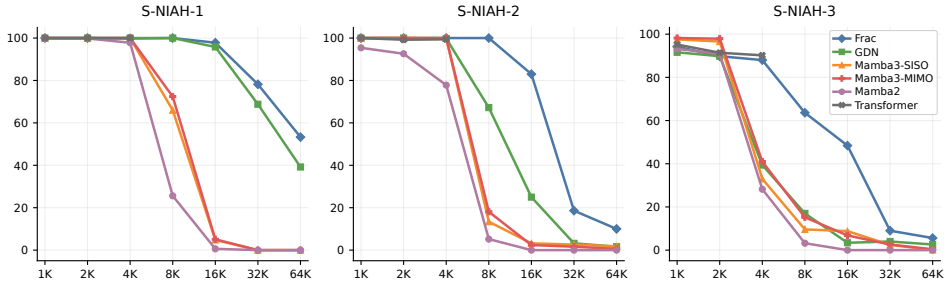
Language models share data and token budgets

- Approximately 1.3B parameters, pretrained from scratch on 100B tokens.
- Deduplicated FineWeb-Edu; Llama-2 tokenizer with a 32K vocabulary.
- Fully packed 4K training sequences; hidden size 2048, 16 heads.
- AdamW, learning rate 6×10^{-4}, 10% warm-up, batch 1M tokens.

Model	Layers
FRAC / Mamba2 / Mamba3-SISO	48
Mamba3-MIMO	44
GDN	22
Transformer (Qwen3 architecture)	26

Paper, Sec. 4.2, p. 7; Appendix E.3, pp. 23–24. Depth varies to match parameter counts.

NIAH improves beyond 4K, but retrieval still degrades



- S-NIAH-1: repetitive filler; S-NIAH-2: number; S-NIAH-3: UUID.
- FRAC's advantage is clear beyond the training range, especially at 8K–16K.
- At 64K, S-NIAH-2/3 accuracy is still around 10% or lower.

Original Fig. 3, PDF p. 8; task definitions, Appendix E.3, p. 24. Last bullet is an approximate visual reading.

LongBench improves by 1.9 score points over GDN

Model	Average score
Transformer	14.7
Mamba2	12.0
GDN	16.0
Mamba3-SISO	14.4
Mamba3-MIMO	11.4
FRAC	**17.9**

- Best on 8 of 14 tasks.
- +1.9 points over GDN; about +11.9% relative using rounded averages.
- Mixed metrics: F1, ROUGE, accuracy, and edit similarity.
- Eight tasks have mean input length below 8K.

Reading the reported gain

The text says "1.9%"; the table supports a 1.9-point difference in benchmark score.

Paper, Table 2, p. 8; task lengths and metrics, Table 7, p. 24.

LongBench gains coexist with a large code-completion loss

Task	Metric	FRAC	GDN	Transformer
GovReport	ROUGE	**13.8**	7.9	7.2
SAMSum	ROUGE	**28.2**	13.2	11.3
NarrativeQA	F1	**13.2**	11.5	8.7
2WikiMQA	F1	12.2	14.8	15.5
LCC	Edit similarity	11.0	18.7	**26.0**
RepoBench-P	Edit similarity	19.5	16.9	25.6

- Strong gains in several summarization and QA tasks.
- LCC loses 15.0 points to the Transformer; its mean input is only 1,235 tokens.
- Longer retention and precise local continuation are different capabilities.

Paper, selected columns of Table 2, PDF p. 8; Table 7, p. 24. All metrics are higher-is-better.

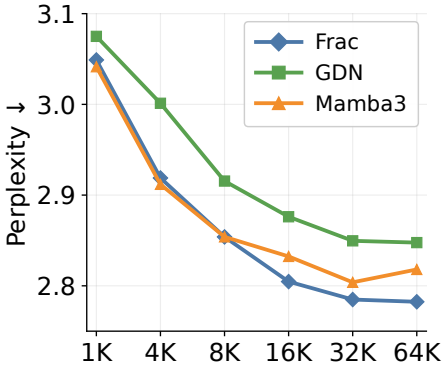
Short-context quality stays competitive

Model	Wiki. PPL↓	Harness avg.↑	Retrieval avg.↑
Transformer	**15.3**	**55.3**	**38.5**
GDN	16.4	54.2	30.9
Mamba3-SISO	16.1	54.6	35.1
Mamba3-MIMO	15.9	**55.3**	36.8
FRAC	16.5	55.1	36.4

- Harness average: eight accuracy metrics, excluding both perplexity columns.
- FRAC trails the best Harness average by 0.2 percentage points.
- Real-world retrieval uses 2K-truncated inputs; FRAC ranks third.

Paper, selected results from Tables 3–4, PDF p. 8; evaluation protocol, p. 24.

DNA modeling benefits from longer training contexts



- HG38; character-level autoregressive modeling.
- About 7M parameters, 8 layers, hidden size 256.
- Lower test perplexity at long contexts, with a plateau near 32K–64K.
- Each model is trained and tested at the same chosen length.

Original Fig. 4, PDF p. 9; Appendix E.4, p. 25. Mamba3 here is the SISO variant.

FRAC leads 16K prefill; Mamba2 trains faster

Model	Prefill 4K	Prefill 16K	Decode	Train 4K
Transformer	**79,173**	56,297	**62**	218,184
Mamba2	41,060	67,571	34	**278,620**
GDN	35,481	58,384	20	215,357
Mamba3-SISO	44,998	74,314	27	238,924
Mamba3-MIMO	30,710	46,707	25	200,994
FRAC	49,946	**77,490**	33	241,598

- All values: tokens/s. Prefill is single-card, batch size 1.
- At 16K: FRAC is 14.7% faster than Mamba2 and 37.6% faster than Transformer.
- Decode averages 64 generated tokens; Transformer uses a 16K prefix.
- Training uses the pretraining setup; FRAC is 13.3% below Mamba2.

Paper, Table 8, Appendix F, PDF p. 25. Accelerator model is unspecified; training uses 16 cards.

The structured prior matters beyond having many states

Variant	ProofPile	PG19	GovReport
FRAC	**47.2**	**28.4**	**11.3**
Fixed $\alpha = 1$	61.3	36.2	19.7
Without write prior	68.5	41.6	18.2
Pure multiscale bank	76.3	47.1	17.2
Without softmax read/write	80.4	36.9	15.5
Freely learned timescales	54.1	34.3	13.5
Without direct feedthrough D	70.9	46.4	23.6

- 390M parameters, 30B tokens, 4K training, 16K evaluation; PPL↓.
- Pure multiscale bank changes α, both priors, and timescale initialization together.

Paper, Table 9 and ablation definitions, Appendix G, PDF pp. 26–27.

Sixteen modes balance fit quality and model cost

Independent homogeneous-kernel fit

M	Mean maximum error
8	1.34×10^{-3}
16	6.82×10^{-4}
32	6.45×10^{-4}

- Fit positive simplex weights separately for each $\alpha \in [0.10, 0.99]$.
- Normalized time $s \in [4, 6553.6]$; geometric $\tau \in [1, 2^{17}]$.
- This measures bank expressivity, not learned token-wise kernel error.

In the LM ablation, $M = 8$ worsens all three PPLs. $M = 32$ improves only GovReport, with about 5% more parameters and 7% slower computation.

Paper, Tables 9–10 and their protocols, Appendix G, PDF pp. 26–27.

The useful interpretation has three boundaries

1. **Finite temporal coverage.** The theorem concerns a finite horizon. Under frozen controls, a finite mixture ultimately has an exponential tail.
2. **Retention versus retrieval.** Better long-range scores coexist with poor LCC results and low accuracy at the hardest NIAH lengths.
3. **Theory versus architecture.** Learned routing, token-varying controls, and the modified write factor extend the fixed-system construction.

Future direction proposed by the authors: combine fractional transitions with delta-rule updates.

Paper, Sec. 6, p. 10; Appendix D, p. 21; Appendix H.1, p. 27. Finite-mixture tail statement: reading analysis.

Memory geometry is a useful design axis for SSMs

Mechanism

Fractional kernels $\rightarrow$ geometric exponential bank $\rightarrow$ selective read/write $\rightarrow$ efficient affine scan.

Evidence

Stronger long-range extrapolation and LongBench scores; short-context quality remains competitive.

Practical lesson

Choose temporal coverage and routing together, and reproduce the actual Algorithm 1 write rule.

Kobyzev et al., Fractional State Space Transition for Long Sequence Modeling, 2026. arXiv:2609.36314v1.

Backup: why the two kernel norms suffice

$$H_{\alpha,\lambda}(\tau) = \frac{R_{\alpha,\lambda}(\tau)}{\lambda\tau}, \qquad c_m = h\tau_m R_{\alpha,\lambda}(\tau_m), \quad d_m = h\tau_m H_{\alpha,\lambda}(\tau_m).$$

Let E_M and g_M denote the finite sums. Then

$$\begin{aligned} |h(t) - \tilde{h}_M(t)| &\le |h(0)|\,|E_\alpha(-\lambda t^\alpha) - E_M(t)| \\ &\quad + \|u\|_\infty \int_0^T |g_{\alpha,\lambda}(r) - g_M(r)|\,dr \\ &\le \varepsilon(|h(0)| + \|u\|_\infty). \end{aligned}$$

- Fixed α, λ; bounded input; finite T; $s_m(0) = h(0)$.
- The impulse kernel is singular at zero but integrable, so L^1 is sufficient.
- The theorem does not supply an explicit guarantee for the chosen $M = 16$.

Paper, Theorem 2 and Proposition 3, pp. 4–5; Appendices B–C, pp. 17–21.

Backup: dimensions and one-token computation

$$u_t \in \mathbb{R}^{H \times d_h}, \quad \Delta_t, \alpha_t, \lambda_t \in \mathbb{R}^H, \quad S_t \in \mathbb{R}^{H \times M \times d_h}.$$

1. Form $a_t = \Delta_t \lambda_t^{1/\alpha_t}$, then $\tilde{\rho}_{t,m} = e^{-a_t/\tau_m}$.
2. Form $c_t = \mathrm{softmax}(-\alpha_t \log \tau + g_{\mathrm{read}}(u_t))$ and independent b_t.
3. Update $S_t = \tilde{\rho}_t \odot S_{t-1} + (1 - \tilde{\rho}_t) b_t \odot u_t$.
4. Read $h_t = \sum_m c_{t,:,m} \odot S_{t,:,m,:}$; mixer adds Du_t.

Mode weights are broadcast over feature coordinates. The content projections use $g(u) = \sigma(p)Wu$, with separate read and write parameters.

Paper, Algorithm 1, Appendix D, PDF p. 21; direct feedthrough, Sec. 3.3, p. 6.

Backup: the effective memory is content dependent

For zero initial state, unrolling the actual recurrence gives

$$h_t = \sum_{k \le t} K_{t,k} u_k, \quad K_{t,k} = \sum_m c_{t,m}(1 - \tilde{\rho}_{k,m}) b_{k,m} \prod_{j=k+1}^{t} \tilde{\rho}_{j,m}.$$

- Current read routing, past write routing, and intervening clocks all matter.
- The selective model generally has no fixed kernel depending only on $t - k$.

Stable within-chunk computation

$$\Gamma_{t,k} = \sum_m c_{t,m} \kappa_{k,m} e^{\ell_{t,m} - \ell_{k,m}}, \qquad \ell_{t,m} = \sum_{j=t_{\text{start}}}^{t} \log \tilde{\rho}_{j,m}.$$

Top equation: teaching derivation from Algorithm 1. Stable implementation: Appendix F, PDF p. 26.