Mamba-3

Improved Sequence Modeling using State Space Principles

Reading presentation

Paper: Aakash Lahoti, Kevin Y. Li, Berlin Chen, Caitlin Wang, Aviv Bick, J. Zico Kolter, Tri Dao, and Albert Gu

arXiv:2603.15569v1 March 16, 2026
Prepared October 20, 2026 25-minute group discussion

Richer state updates at a competitive decoding cost.

arxiv.org/abs/2603.15569v1

Make each fixed-state update do more useful work

- Fixed recurrent state avoids a growing KV cache.
- Simple positive decay limits state-transition behavior.
- Low-FLOP decoding can still leave GPU compute idle.

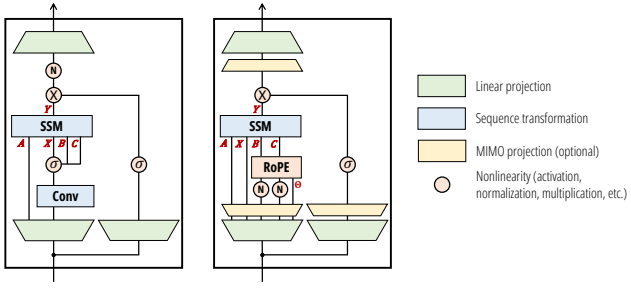
Design change	What changes	Main objective
Exponential-trapezoidal	Input integration	Better local mixing
Complex transitions	State rotation	State tracking
MIMO	Read/write rank	Quality per decode cost

Interpretation

Optimize the content of an update as well as its memory footprint.

Paper, Secs. 1 and 3, PDF pp. 1–2, 4–11.

Mamba-3 changes the recurrence and its surrounding block



- BC normalization and learned biases.
- Data-dependent rotary embeddings.
- Optional MIMO projections.
- External short convolution removed.

Original Fig. 2; Sec. 3.4, PDF pp. 11–12. Blocks alternate with pre-norm SwiGLU.

Mamba-2 uses positive scalar decay and rank-one writes

$$\mathbf{h}_t = \alpha_t \mathbf{h}_{t-1} + \Delta_t \mathbf{B}_t x_t, \qquad y_t = \mathbf{C}_t^\top \mathbf{h}_t$$
$$\alpha_t = e^{\Delta_t A_t} \in (0,1), \qquad A_t < 0, \quad \Delta_t > 0$$

- N: real state coordinates. P: channels per head.
- Stack P channels: $\mathbf{H}_t \in \mathbb{R}^{N \times P}$ and $\Delta_t \mathbf{B}_t \mathbf{x}_t^\top$ is a rank-one write.
- A larger Δ_t increases forgetting and current-input weight.

Interpretation

The update can scale memory, but it cannot directly rotate it.

Paper, Sec. 2.2, Eq. (1), PDF p. 3; Sec. 3.3, p. 9.

Discretization separates decay from input integration

Freeze the scalar $A(t)$ at the right endpoint A_t:

$$\mathbf{h}_t \approx e^{\Delta_t A_t}\mathbf{h}_{t-1} + \int_{\tau_{t-1}}^{\tau_t} e^{(\tau_t-\tau)A_t}\mathbf{B}(\tau)x(\tau)\,d\tau.$$

Right endpoint: exponential-Euler

$$\text{input term} = \Delta_t \mathbf{B}_t x_t$$

Recovers the Mamba-1/2 implementation.

Two endpoints: exponential-trapezoidal

$$\text{input term} = \beta_t \mathbf{B}_{t-1} x_{t-1} + \gamma_t \mathbf{B}_t x_t$$

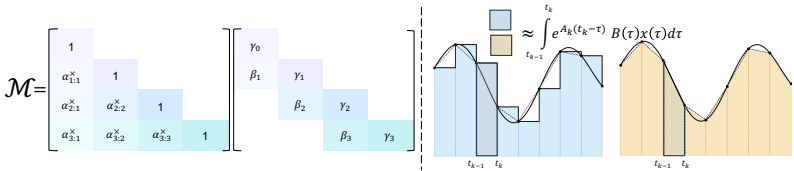
Learns how to mix adjacent state inputs.

Paper, Sec. 3.1.1, Eqs. (4)–(6), PDF pp. 5–6; App. A, p. 22.

The new recurrence contains a two-tap state-input filter

$$\boxed{\mathbf{h}_t = \alpha_t \mathbf{h}_{t-1} + \beta_t \mathbf{B}_{t-1} x_{t-1} + \gamma_t \mathbf{B}_t x_t}$$

$$\alpha_t = e^{\Delta_t A_t}, \quad \beta_t = (1 - \lambda_t)\Delta_t \alpha_t, \quad \gamma_t = \lambda_t \Delta_t$$



- The filter acts on $\mathbf{B}_t x_t$ inside the recurrence.
- SSD mask = decay matrix $\times$ two-band matrix.

Paper, Fig. 1, PDF p. 4; Secs. 3.1.2–3.1.3, Eq. (7), pp. 6–7.

Higher quadrature order requires a condition on the gate

$$\lambda_t = \tfrac{1}{2} + O(\Delta_t) \quad \Longrightarrow \quad \text{input-integral error} = O(\Delta_t^3)$$

- Requires the smoothness and stability assumptions in Appendix A.
- Freezing a varying $A(t)$ already introduces $O(\Delta_t^2)$ local error.
- The default $\lambda_t = \sigma(u_t)$ does not enforce the condition.

Gate, 440M component ablation	Perplexity ↓
Learned $\sigma(u_t)$	**15.72**
Fixed $1/2$	15.76
Fixed 1 (Euler)	15.81

Paper, Remark 3, PDF p. 6; App. A.2–A.3, pp. 23–24, Table 8.

Complex transitions add input-dependent rotations

$$e^{\Delta_t(A_t+i\theta_t)} = e^{\Delta_t A_t}\left(\cos(\Delta_t\theta_t) + i\sin(\Delta_t\theta_t)\right)$$

$$\mathbf{h}_t = \alpha_t \mathbf{R}_t \mathbf{h}_{t-1} + \Delta_t \mathbf{B}_t x_t, \qquad R(\phi) = \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix}$$

- $N/2$ complex coordinates use N real coordinates.
- Amplitude decay and phase evolution become separate controls.
- Teaching example: rotate by πx_t to toggle a parity state.

Interpretation

The restricted positive transition is the obstacle. Negative real transitions can also solve parity.

Paper, Prop. 2, Eqs. (8)–(9), PDF pp. 7–8; Table 5b, p. 14. Last point: algebraic interpretation.

A rotating coordinate system keeps the core update scalar

$$\mathbf{Q}_t = \prod_{s=0}^{t} \mathbf{R}_s^\top, \quad \overline{\mathbf{B}}_t = \mathbf{Q}_t \mathbf{B}_t, \quad \overline{\mathbf{C}}_t = \mathbf{Q}_t \mathbf{C}_t, \quad \widetilde{\mathbf{h}}_t = \mathbf{Q}_t \mathbf{h}_t$$

$$\boxed{\widetilde{\mathbf{h}}_t = \alpha_t \widetilde{\mathbf{h}}_{t-1} + \beta_t \overline{\mathbf{B}}_{t-1} x_{t-1} + \gamma_t \overline{\mathbf{B}}_t x_t} \qquad y_t = \overline{\mathbf{C}}_t^\top \widetilde{\mathbf{h}}_t$$

- Per-pair phase: $\phi_t[j] = \sum_{s=0}^{t} \Delta_s \theta_s[j]$.
- Rotate projections with $R(-\phi_t[j])$; use scalar-state kernels.
- Unlike standard RoPE, phase increments depend on input content.
- The previous input uses $\mathbf{Q}_{t-1}$; the current input uses $\mathbf{Q}_t$.

Paper, Props. 3–4, Eqs. (10)–(11), PDF p. 8; App. B, pp. 25–26.

MIMO raises read/write rank while keeping the same state

Per-head tensor	SISO	MIMO
$\mathbf{B}_t, \mathbf{C}_t$	N	$N \times R$
Input	P	$P \times R$
Persistent state $\mathbf{H}_t$	$N \times P$	$N \times P$
Write $\mathbf{B}_t \mathbf{X}_t^\top$	Outer product	Matrix product

$$\mathbf{H}_t = \alpha_t \mathbf{H}_{t-1} + \Delta_t \mathbf{B}_t \mathbf{X}_t^\top, \qquad \mathbf{Y}_t = \mathbf{H}_t^\top \mathbf{C}_t \in \mathbb{R}^{P \times R}$$

- For $R \ll N, P$, compute grows while state traffic remains dominant.
- Arithmetic intensity grows roughly with R; matmuls use tensor cores.

Paper, Sec. 3.3, Table 2, PDF pp. 9–10. Two-term form shown; the trapezoidal extension is analogous.

MIMO trades MLP width for richer state updates

$$\mathbf{x}'_t = W_{X'}\mathbf{U}_t, \qquad X_t[p,r] = W_X[p,r]\,x'_t[p]$$

- Factorized per-head input projection: $DP + PR$ parameters instead of DPR.
- Similar rank expansion/compression for gates and outputs.
- At 1.5B, MLP width changes from 4096 to 3824.

Training cost remains higher

A naive R^2 SISO expansion becomes approximately R times the dominant FLOPs with shorter chunks:

$$C_{\text{chunk}}R = N = P: \qquad 8TN^2 \longrightarrow 8TRN^2.$$

Paper, Sec. 3.3, PDF p. 10; App. C, p. 27. Authors report about 2x kernel slowdown at rank 4.

Different experiments use different training budgets

Experiment	Configuration	Comparison target
Main language modeling	180M–1.5B, 100B tokens	General quality
Component ablations	440M, ∼8.8B tokens	Mechanism contribution
State-size sweep	440M, ∼17.6B tokens	Capacity–quality tradeoff
State tracking	1/3 layers; test length 256	Algorithmic behavior
Decode kernels	H100, batch 128	Actual latency

- Main models: FineWeb-Edu, 2K context, Llama-3.1 tokenizer, BF16.
- Default Mamba settings: $N = 128$, $P = 64$; MIMO rank $R = 4$.
- Main baselines: Transformer, Mamba-2, and Gated DeltaNet.

Paper, Sec. 4, PDF pp. 12–15; Apps. D–G, pp. 27–31. Approximate token counts computed from reported ratios.

At 1.5B, MIMO adds 1.2 points over Mamba-3 SISO

Model	FineWeb-Edu ppl ↓	Seven-task average ↑
Transformer	10.51	55.4
Gated DeltaNet	10.45	55.8
Mamba-2	10.47	55.7
Mamba-3 SISO	10.35	56.4
Mamba-3 MIMO	**10.24**	**57.6**

- SISO vs. GDN: +0.6 percentage points.
- MIMO vs. GDN: +1.8 points; vs. Mamba-2: +1.9 points.
- Matched parameters and data; MIMO uses more training compute.

Paper, Table 3, PDF p. 12. Accuracy average excludes perplexities; HellaSwag and ARC-C use normalized accuracy.

Fixed-state models still lose information needed for retrieval

Pure 1.5B model	SWDE	SQuAD	FDA	TriviaQA
Transformer	48.9	46.6	58.4	67.5
GDN	32.7	40.0	28.3	63.5
Mamba-2	30.7	39.1	23.7	64.3
Mamba-3 SISO	28.5	40.1	23.4	64.5
Mamba-3 MIMO	36.3	41.7	29.3	64.5

Cloze-formatted tasks, truncated to 2K; higher scores are better.

Interpretation

MIMO improves several retrieval scores, while semi-structured and unstructured information extraction remains difficult.

Paper, Sec. 4.1.2 and Table 4, PDF pp. 13–14.

Length-generalization gains depend on the retrieval task

Pure 1.5B model, evaluated at 4K	NIAH-1	NIAH-2	NIAH-3
Transformer, standard RoPE	0.0	0.0	0.0
GDN	99.8	49.8	34.2
Mamba-2	62.0	11.8	13.4
Mamba-3 SISO	88.2	50.6	34.2
Mamba-3 MIMO	93.0	40.4	25.6

- Training context is 2K; 4K tests extrapolation.
- SISO improves all three tasks over Mamba-2.
- MIMO improves NIAH-1 but reduces NIAH-2/3 scores.
- Transformer failure is specific to this RoPE/extrapolation setup.

Paper, Table 4, PDF p. 14. NIAH-1/2/3 denote the three NIAH-Single task variants.

Hybrid retrieval is sensitive to the extra output norm

Five linear layers per one NoPE attention layer; all variants keep BCNorm.

Hybrid Mamba-3 SISO	FDA, 2K	NIAH-1, 4K	NIAH-2, 4K	NIAH-3, 4K
No extra norm	65.9	36.2	9.4	8.8
Pre-gate grouped RMS	52.4	100.0	96.0	56.8
Pre-gate default RMS	67.3	86.2	97.8	90.2

- Hybrid FDA score without extra norm: 65.9, vs. pure SISO's 23.4.
- Grouped pre-gate norm improves 4K retrieval, but hurts FDA.
- The default pre-gate norm wins on other task/length combinations.

Interpretation

Norm type and placement are part of the retrieval tradeoff.

Paper, Sec. 4.1.2, PDF p. 13; Table 4, p. 14; Table 9, p. 28.

Biases and trapezoidal writes make short convolution optional

440M SISO variant	Test perplexity ↓
No BC bias, no trapezoidal update	16.68
Remove BC bias	16.49
Full Mamba-3	**15.72**
Full Mamba-3 + short convolution	15.85

- Trapezoidal writes mix adjacent state inputs.
- BC biases add data-independent components to the projections.
- Reintroducing the external convolution does not help this setup.

Paper, Table 5a, PDF p. 14; mechanism discussion in Sec. 3.4, pp. 11–12. Chinchilla-ratio training.

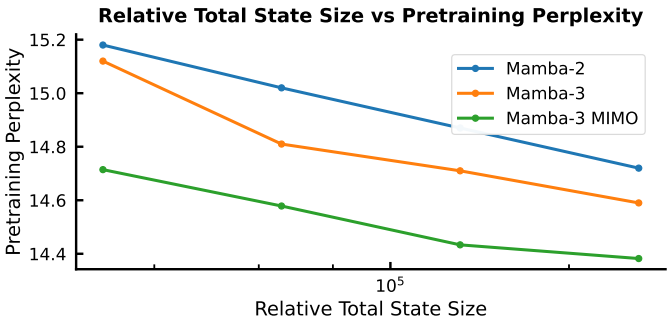
Content-dependent phase enables state tracking

Model	Parity	Arithmetic no brackets	Arithmetic with brackets
Mamba-3	100.00	98.51	87.75
Mamba-3, standard RoPE	1.56	20.70	2.62
Mamba-3, no RoPE	2.27	1.49	0.72
Mamba-2	0.90	47.81	0.88
GDN, eigenvalues in $[-1, 1]$	100.00	99.25	93.50

- Scaled accuracy (%), higher is better; not raw accuracy.
- Train up to length 160; evaluate at 256.
- One layer for parity, three for arithmetic; hyperparameter sweep.

Paper, Table 5b, PDF p. 14; App. D, p. 28. Mamba-2 has partial success on unbracketed arithmetic.

Mamba-3 uses a given state capacity more effectively



- 440M models, about 17.6B training tokens.
- $N \in \{16, 32, 64, 128\}$.
- Lower curves: better quality at the same state size.
- Horizontal axis is a speed proxy, not measured latency.

Original Fig. 3; Sec. 4.3, PDF pp. 14–15. Half the state size does not establish half the end-to-end latency.

BF16 decoding is fast; MIMO adds a modest cost over SISO

	FP32		BF16	
Model	$N = 64$	$N = 128$	$N = 64$	$N = 128$
Mamba-2	0.295	0.409	0.127	0.203
GDN	0.344	0.423	0.176	0.257
Mamba-3 SISO	0.310	0.399	0.110	0.156
Mamba-3 MIMO	0.333	0.431	0.137	0.179

Single-step kernel latency in **ms**; H100, batch 128, 1.5B configuration.

- BF16, $N = 128$: MIMO latency is 11.8% lower than Mamba-2 and 14.7% higher than SISO.
- FP32, $N = 64$: Mamba-2 is fastest.
- Baseline Triton vs. Mamba-3 CuTe/Triton implementations.

Paper, Table 6, PDF p. 15; App. G.1, pp. 30–31. Relative changes calculated from reported times.

Prefill and full generation reveal the MIMO overhead

Model	2K prefill	2K total	16K total
vLLM, Llama-3.2-1B	1.08	20.37	976.50*
GDN, 1.5B	2.01	18.22	145.87
Mamba-2, 1.5B	2.02	18.62	149.02
Mamba-3 SISO, 1.5B	2.02	17.57	140.61
Mamba-3 MIMO, 1.5B	2.42	18.96	151.81

Times in **seconds**; "total" means prefill L + decode L tokens.

- H100-SXM 80GB, target batch 128, three repetitions.
- *16K vLLM: batch 16 time multiplied by 8 because of OOM.
- Different model sizes and kernels; compare the measured setting.

Paper, Table 7, PDF p. 15; App. G.2, pp. 30–31. Recurrent models use state passing when needed at 16K.

Three ways to improve the value of a recurrent state

1. **Improve the write.**
   Two-endpoint integration and BC biases provide effective local mixing.
2. **Improve the transition.**
   Input-dependent rotations support state tracking through efficient transformed projections.
3. **Improve compute per state access.**
   MIMO increases modeling power at a competitive decode cost, with higher training and prefill costs.

Interpretation

State capacity, transition structure, and per-step compute are separate architecture choices.

Paper, Sec. 6, PDF p. 18. Evidence: up to 1.5B parameters and 100B training tokens; retrieval remains task-dependent.

Backup: the input-integral error exposes the gate condition

Let $g(\tau) = e^{(\tau_t - \tau)A_t}\mathbf{B}(\tau)x(\tau)$ and

$$Q_\lambda = \Delta_t\left[(1-\lambda_t)g(\tau_{t-1}) + \lambda_t g(\tau_t)\right].$$

Taylor expansion around the left endpoint gives

$$\int_{\tau_{t-1}}^{\tau_t} g(\tau)\,d\tau - Q_\lambda = \left(\tfrac{1}{2} - \lambda_t\right)\Delta_t^2 g' + \left(\tfrac{1}{6} - \tfrac{\lambda_t}{2}\right)\Delta_t^3 g'' + O(\Delta_t^4).$$

- $\lambda_t = \frac{1}{2} + O(\Delta_t)$ suppresses the leading term.
- This concerns quadrature after the approximation of $A(t)$.
- Learned model quality need not track numerical integration order.

Paper, App. A.2–A.3, PDF pp. 23–24; Table 8. Derivatives above are evaluated at the left endpoint.

Backup: the SSD mask follows by unrolling three steps

With $\mathbf{v}_t = \mathbf{B}_t x_t$:

$$\begin{aligned}
\mathbf{h}_0 &= \gamma_0 \mathbf{v}_0, \\
\mathbf{h}_1 &= (\alpha_1 \gamma_0 + \beta_1)\mathbf{v}_0 + \gamma_1 \mathbf{v}_1, \\
\mathbf{h}_2 &= \alpha_2(\alpha_1 \gamma_0 + \beta_1)\mathbf{v}_0 + (\alpha_2 \gamma_1 + \beta_2)\mathbf{v}_1 + \gamma_2 \mathbf{v}_2.
\end{aligned}$$

$$L_{ts} = \left(\prod_{k=s+2}^{t} \alpha_k \right) (\alpha_{s+1}\gamma_s + \beta_{s+1}) \quad (t > s), \qquad L_{tt} = \gamma_t.$$

$$\mathbf{Y} = (\mathbf{L} \odot \mathbf{C}\mathbf{B}^\top)\mathbf{X}, \qquad \mathbf{L} = \mathbf{L}_{\text{decay}} \mathbf{L}_{\text{2band}}.$$

Paper, Sec. 3.1.3, Eq. (7), PDF pp. 6–7; App. A.1, pp. 22–23.

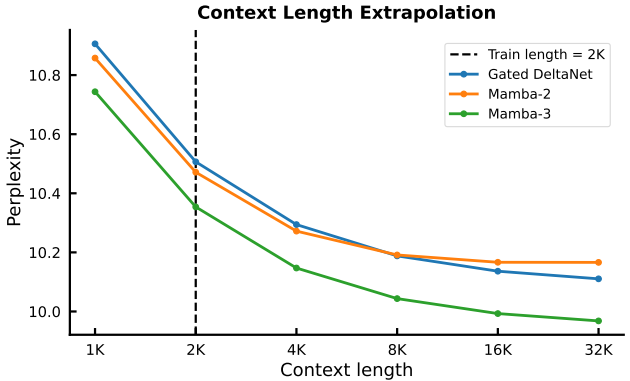
Backup: quality gains vary across model scales

Average accuracy (%)	180M	440M	880M	1.5B
Transformer	42.8	49.6	52.8	55.4
GDN	43.2	49.7	53.7	55.8
Mamba-2	42.9	49.6	53.4	55.7
Mamba-3 SISO	43.4	49.8	54.4	56.4
Mamba-3 MIMO	43.5	51.0	55.3	57.6

- SISO exceeds prior baselines at each scale, sometimes by a small margin.
- It does not lead every metric: at 180M, FineWeb-Edu ppl is 16.59 vs. GDN's 16.52.
- All models use 100B tokens; model scales have different tokens/parameter ratios.

Paper, Table 3, PDF p. 12.

Backup: language-model perplexity improves beyond 2K



Mamba-3 continues to benefit from longer context; Mamba-2 levels off.

Original Fig. 4, PDF p. 28. 1.5B pretrained models, held-out FineWeb-Edu; dashed line marks the training length.

Backup: positive initialization and both BC biases help

Bias initialization	Learned	ppl ↓
All ones	Yes	15.72
All zeros	Yes	16.57
All ones	No	15.80
$\mathcal{U}(0,1)$	Yes	15.76
$\mathcal{U}(-1,1)$	Yes	16.07

B bias	C bias	ppl ↓
No	No	16.52
Yes	No	16.68
No	Yes	15.98
Yes	Yes	15.69

- Biases are head-specific, channel-wise, and added after BCNorm.
- These are separate ablation groups; retain each group's reported values.

Paper, App. F, Tables 10a–10b, PDF pp. 29–30. 440M, Chinchilla-ratio training.

Backup: a unified MIMO update combines all three changes

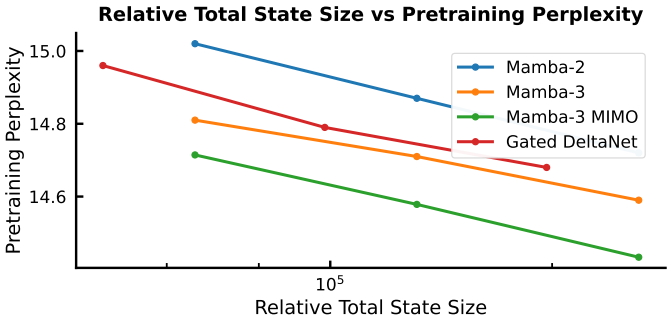
$$\widetilde{\mathbf{H}}_t = \alpha_t \widetilde{\mathbf{H}}_{t-1} + \beta_t \overline{\mathbf{B}}_{t-1} \mathbf{X}_{t-1}^\top + \gamma_t \overline{\mathbf{B}}_t \mathbf{X}_t^\top,$$
$$\mathbf{Y}_t = \widetilde{\mathbf{H}}_t^\top \overline{\mathbf{C}}_t \in \mathbb{R}^{P \times R},$$
$$\mathbf{Y}'_t = \mathbf{Y}_t \odot \mathrm{SiLU}(\mathbf{Z}_t).$$

$$O'_t[p] = \sum_{r=1}^{R} W_{O'}[p, r] Y'_t[p, r], \qquad \mathbf{O}_t = W_O \mathbf{O}'_t.$$

- Rotated $\overline{\mathbf{B}}, \overline{\mathbf{C}}$ each have shape $N \times R$.
- The persistent state keeps shape $N \times P$.
- Prior-input caches and cumulative phases are additional small state.

Teaching synthesis of Prop. 4, Sec. 3.3, and App. C; PDF pp. 8–11, 27.

Backup: compare total state size across head structures



- GDN: multi-head; Mamba: multi-value.
- Equal nominal dimensions need not mean equal total state.
- GDN keeps $v_{\text{expand}} = 2$ and varies head dimension.
- All models parameter-matched at 440M.

Original Fig. 6; App. E, PDF p. 29. Primary reference: Lahoti et al., arXiv:2603.15569v1.