

# MuonSSM: Orthogonalizing State Space Models for Sequence Modeling

Thai-Khanh Nguyen<sup>1,3\*</sup>

**Ngoc-Bich-Uyen Vo**<sup>2\*</sup>

Thieu N. Vo<sup>4†</sup>

Tan M. Nguyen<sup>5†</sup>

Cuong Pham<sup>2†</sup>



<sup>1</sup> Dainam University

<sup>2</sup> Posts and Telecommunications Institute of Technology

<sup>3</sup> Hanoi University of Science and Technology

<sup>4</sup> University of Bath

<sup>5</sup> National University of Singapore

*\*Equal contribution*

*†Co-last authors*

# State Space Models as Associative Memory

At each step the model **maintains** a memory matrix  $\mathbf{S}_t$  and **writes** a key-value pair via a first-order affine update<sup>(1)</sup>:

Updated memory state:

$$\mathbf{S}_t = \mathbf{S}_{t-1} \left( \alpha_t (\mathbf{I}_m - \beta_t \eta \mathbf{k}_t \mathbf{k}_t^\top) \right) + \beta_t \mathbf{v}_t \mathbf{k}_t^\top, \quad (1)$$

$\mathbb{R}^{d \times m}$        $(0, 1]$        $> 0$        $\mathbb{R}^d \mathbb{R}^m$

memory retention      update magnitude

recall correction

(1) Behrouz, Ali, et al., *It's All Connected: A Journey Through Test-Time Memorization, Attentional Bias, Retention, and Online Optimization (ICLR, 2026)*.

# State Space Models as Associative Memory

Updated memory state: 
$$\mathbf{S}_t = \mathbf{S}_{t-1} \left( \alpha_t (\mathbf{I}_m - \beta_t \eta \mathbf{k}_t \mathbf{k}_t^\top) \right) + \beta_t \mathbf{v}_t \mathbf{k}_t^\top, \quad (1)$$

$\mathbb{R}^{d \times m}$        $(0, 1]$        $> 0$        $\mathbb{R}^d \mathbb{R}^m$   
 memory retention      update magnitude  
 recall correction

Model	$\alpha_t$	$\beta_t$	$\eta$	Update Rule
Mamba (Gu & Dao, 2024)	$\alpha_t$	1	0	$\mathbf{S}_t = \alpha_t \mathbf{S}_{t-1} + \mathbf{v}_t \mathbf{k}_t^\top$
DeltaNet (Yang et al., 2024b)	1	$\beta_t$	1	$\mathbf{S}_t = \mathbf{S}_{t-1} (\mathbf{I} - \beta_t \mathbf{k}_t \mathbf{k}_t^\top) + \beta_t \mathbf{v}_t \mathbf{k}_t^\top$
Gated DeltaNet (Yang et al., 2024a)	$\alpha_t$	$\beta_t$	1	$\mathbf{S}_t = \mathbf{S}_{t-1} (\alpha_t (\mathbf{I} - \beta_t \mathbf{k}_t \mathbf{k}_t^\top)) + \beta_t \mathbf{v}_t \mathbf{k}_t^\top$
LongHorn (Liu et al., 2024)	1	$\frac{\beta_t}{1 + \beta_t \mathbf{k}_t^\top \mathbf{k}_t}$	1	$\mathbf{S}_t = \mathbf{S}_{t-1} \left( \mathbf{I} - \frac{\beta_t}{1 + \beta_t \mathbf{k}_t^\top \mathbf{k}_t} \mathbf{k}_t \mathbf{k}_t^\top \right) + \frac{\beta_t}{1 + \beta_t \mathbf{k}_t^\top \mathbf{k}_t} \mathbf{v}_t \mathbf{k}_t^\top$

**Table 1.** Special cases of recent SSMs recovered from the general associative memory update in Eq. (1)

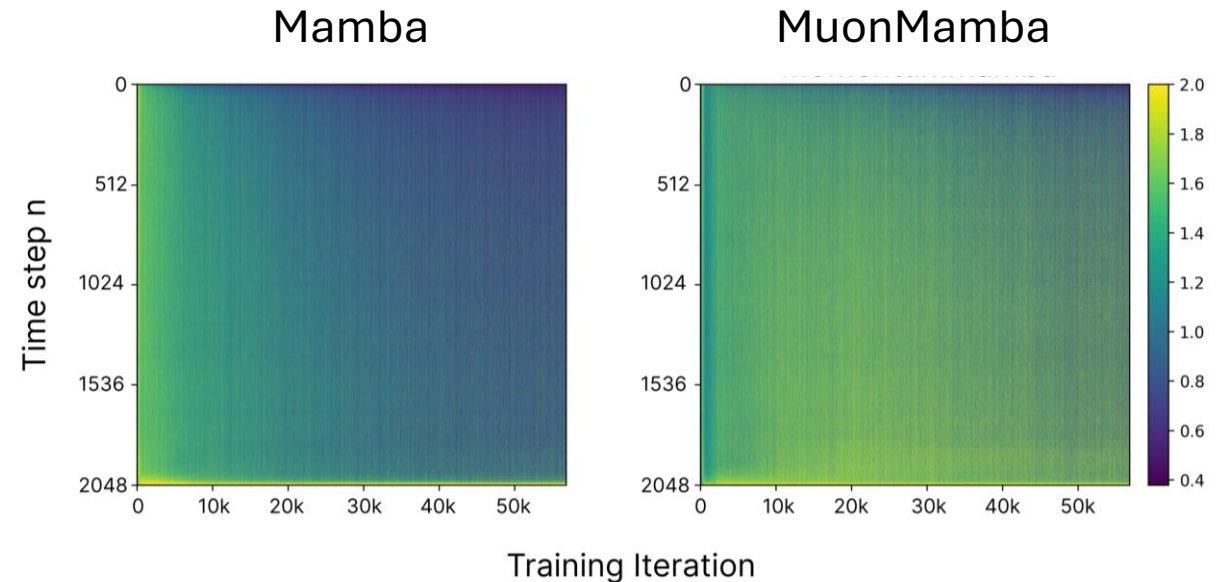
# The Problem: First-Order Updates Degrade

The change in memory of all SSMs:  $\Delta \mathbf{S}_t \propto (v_t - \alpha_t \eta \mathbf{S}_{t-1} k_t) k_t^T$  (first-order rank-1 updates)

## Limitations:

- **Spectral anisotropy:** Singular values become highly non-uniform.
- **Memory interference:** New updates can overwrite previous information
- **Gradient degradation:** Vanishing gradients through  $\prod_{n=T}^t \mathbf{D}_n$  where  $\mathbf{D}_n = \alpha_n (\mathbf{I}_m - \beta_n \eta k_n k_n^T)$ .

→ Current SSMs (gating, normalization, transition parameterization) only address this **indirectly**.



**Figure 1.** MuonMamba (our method) shows more uniform gradient flow over long contexts than Mamba.

# MuonSSM: Orthogonalizing State Space Models

$$\mathbf{S}_t = \mathbf{S}_{t-1} \left( \alpha_t (\mathbf{I}_m - \beta_t \eta \mathbf{k}_t \mathbf{k}_t^T) \right) + \beta_t \mathbf{v}_t \mathbf{k}_t^T \quad (1)$$

- **NS**( $\cdot$ ) – single-iteration Newton-Schulz
  - $\tau > 0$  – update scaling factor

MuonSSM

$$\mathbf{NS}(\tau \beta_t \mathbf{v}_t \mathbf{k}_t^T)$$

# Single-iteration Newton-Schulz

**Definition. Single-iteration Newton–Schulz (NS).**

For  $\mathbf{X} \in \mathbb{R}^{d \times m}$ , define:

$$\tilde{\mathbf{X}} = \frac{\mathbf{X}}{\max(\|\mathbf{X}\|_F, \delta)}$$
$$\text{NS}(\mathbf{X}) = \left( a + b\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T + c(\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T)^2 \right) \tilde{\mathbf{X}}$$

$(a, b, c) = (3.4445, -4.7750, 2.0315)$  (Jordan et al.<sup>(2)</sup>) and  $\delta > 0$ .

## ***MuonOptimizer***<sup>(2)</sup>

**Newton-Schulz iteration** is to approximately orthogonalize the update matrix.

*“orthogonalization effectively increases the scale of other **“rare directions”** which have **small magnitude** in the update but are nevertheless **important** for learning.”*

(2) Jordan, K. et al. *Muon: An optimizer for hidden layers in neural networks*, 2024

# MuonSSM: Orthogonalizing State Space Models

$$\mathbf{S}_t = \mathbf{S}_{t-1} \left( \alpha_t (\mathbf{I}_m - \beta_t \eta \mathbf{k}_t \mathbf{k}_t^T) \right) + \beta_t \mathbf{v}_t \mathbf{k}_t^T \quad (1)$$

- **NS**( $\cdot$ ) – single-iteration Newton-Schulz
  - $\tau > 0$  – update scaling factor
- $\mathbf{M}_t \in \mathbb{R}^{d \times m}$  – auxiliary momentum matrix
  - $\gamma \in (0,1]$  – momentum decay coefficient

MuonSSM

$$\mathbf{M}_t = \gamma \mathbf{M}_{t-1} + \mathbf{NS}(\tau \beta_t \mathbf{v}_t \mathbf{k}_t^T)$$

$$\mathbf{S}_t = \mathbf{S}_{t-1} \left( \alpha_t (\mathbf{I}_m - \beta_t \eta \mathbf{k}_t \mathbf{k}_t^T) \right) + \mathbf{M}_t$$



# Parallelizability of MuonSSM

## Block-Affine Recurrence

$$\begin{aligned}\mathbf{M}_t &= \gamma \mathbf{M}_{t-1} + \mathbf{NS}(\tau \beta_t \mathbf{v}_t \mathbf{k}_t^T) \\ \mathbf{S}_t &= \mathbf{S}_{t-1} \left( \alpha_t (\mathbf{I}_m - \beta_t \eta \mathbf{k}_t \mathbf{k}_t^T) \right) + \mathbf{M}_t\end{aligned}$$

$$\mathbf{z}_t = [\mathbf{S}_t \quad \mathbf{M}_t] \in \mathbb{R}^{d \times 2m}; \quad \Phi_t = \begin{bmatrix} \alpha_t (\mathbf{I}_m - \beta_t \eta \mathbf{k}_t \mathbf{k}_t^T) & \mathbf{0} \\ \gamma \mathbf{I} & \gamma \mathbf{I} \end{bmatrix}; \quad \Psi_t = \begin{bmatrix} \mathbf{NS}(\tau \beta_t \mathbf{v}_t \mathbf{k}_t^T) \\ \mathbf{NS}(\tau \beta_t \mathbf{v}_t \mathbf{k}_t^T) \end{bmatrix}$$

$$\mathbf{z}_t = \mathbf{z}_{t-1} \Phi_t + \Psi_t \quad (\text{linear recurrence})$$

The recurrence admits an **associative** scan operator.

**Parallel prefix scan** reduces **depth** from  $O(L)$  to  $O(\log L)$  for associative operations, without increasing **total work** beyond  $O(L)$ .

# Theoretical Analysis – Gradient Stability

## Gradient Propagation in MuonSSM

$$\mathbf{Z}_t = \mathbf{Z}_{t-1} \Phi_t + \Psi_t$$

$$\Phi_t = \begin{bmatrix} \mathbf{D}_t & \mathbf{0} \\ \gamma \mathbf{I} & \gamma \mathbf{I} \end{bmatrix}; \quad \mathbf{D}_t = \alpha_t (\mathbf{I}_m - \beta_t \eta \mathbf{k}_t \mathbf{k}_t^T)$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{Z}_{t-1}} = \frac{\partial \mathcal{L}}{\partial \mathbf{Z}_T} \prod_{n=T}^t \Phi_n^T = \frac{\partial \mathcal{L}}{\partial \mathbf{Z}_T} \begin{bmatrix} \prod_{n=T}^t \mathbf{D}_n^T & \sum_{k=t}^T \left( \prod_{j=T}^{k+1} \mathbf{D}_j^T \right) (\gamma \mathbf{I}_m)^{k-t+1} \\ \mathbf{0} & (\gamma \mathbf{I}_m)^{T-t+1} \end{bmatrix}$$

When  $\gamma \approx 1$ , the momentum pathway  $(\gamma \mathbf{I}_m)^{T-t+1}$  preserves gradients even over long horizons.

# Theoretical Analysis – Rank Enrichment

## *Forward Spectral Conditioning of Updates*

**NS** maps  $\sigma \in [0,1]$  through  $\rho(\sigma)$ , ensuring bounded singular values:  $\sigma_{max}(\mathbf{NS}(\mathbf{X})) \leq 1 + \varepsilon_u$

## *Backward Geometry of Newton-Schulz Normalization*

For rank-one  $\mathbf{X}_0 = uw^T$ , the NS Jacobian decomposes into four orthogonal eigenspaces.

$$\lambda = \begin{cases} 0 & \text{direction } uw^T \\ a + b + c & \text{direction } u_{\perp}w^T, uw_{\perp}^T \\ a & \text{direction } u_{\perp}w_{\perp}^T \end{cases}$$

- **NS** amplifies orthogonal gradient directions
  - **Momentum** accumulates less collinear updates
- Increase the effective rank

# Experimental Setup

- **Baselines:** Mamba, LongHorn, Gated DeltaNet – each tested with and without the MuonSSM update mechanism.
- **Controlled comparison:** Identical parameter counts, architectural hyperparameters, and training budgets across all runs – only the memory update mechanism varies.
- **Hardware:** 4 × NVIDIA H100 GPUs.
- **Three modalities:**

Modality	Task	Dataset
Language	Zero-shot reasoning, Long-context retrieval	FineWeb-Edu 10B, S-NIAH
Vision	Classification, Detection, Segmentation	ImageNet, MS COCO, ADE20K
Time-series	Human activity recognition	MuWiGes, UESTC-MMEA-CL, MMAAct

# Language Modeling and Long-Context Retrieval

Architecture	Memory Algorithm	Wiki. ppl ↓	LMB. ppl ↓	LMB. acc ↑	PIQA acc ↑	Hella. acc_n ↑	Wino. acc ↑	ARC-e acc ↑	ARC-c acc_n ↑	SIQA acc ↑	BoolQ acc ↑	Avg. acc ↑
Mamba	Original	42.17	102.51	20.57	63.81	30.10	51.11	52.39	21.75	37.41	59.87	42.13
	+ Muon (Ours)	<b>40.83</b>	<b>89.17</b>	<b>22.84</b>	63.47	<b>33.19</b>	<b>53.36</b>	<b>53.21</b>	<b>25.58</b>	<b>38.33</b>	<b>63.82</b>	<b>44.23</b>
LongHorn	Original	43.06	96.80	22.16	62.79	29.87	52.25	50.34	21.24	36.18	55.02	41.23
	+ Muon (Ours)	<b>41.71</b>	<b>80.98</b>	<b>24.00</b>	62.02	<b>32.85</b>	<b>54.38</b>	<b>51.17</b>	<b>25.12</b>	<b>37.15</b>	<b>59.44</b>	<b>43.27</b>
Gated DeltaNet	Original	39.58	97.92	21.36	62.45	30.02	51.30	51.85	21.42	36.90	55.23	41.32
	+ Muon (Ours)	<b>38.12</b>	<b>83.47</b>	<b>23.91</b>	<b>62.88</b>	<b>33.51</b>	<b>53.76</b>	<b>52.74</b>	<b>23.13</b>	<b>38.02</b>	<b>56.85</b>	<b>43.10</b>

**Table.** Zero-shot reasoning results after pre-training on FineWeb-Edu 10B.

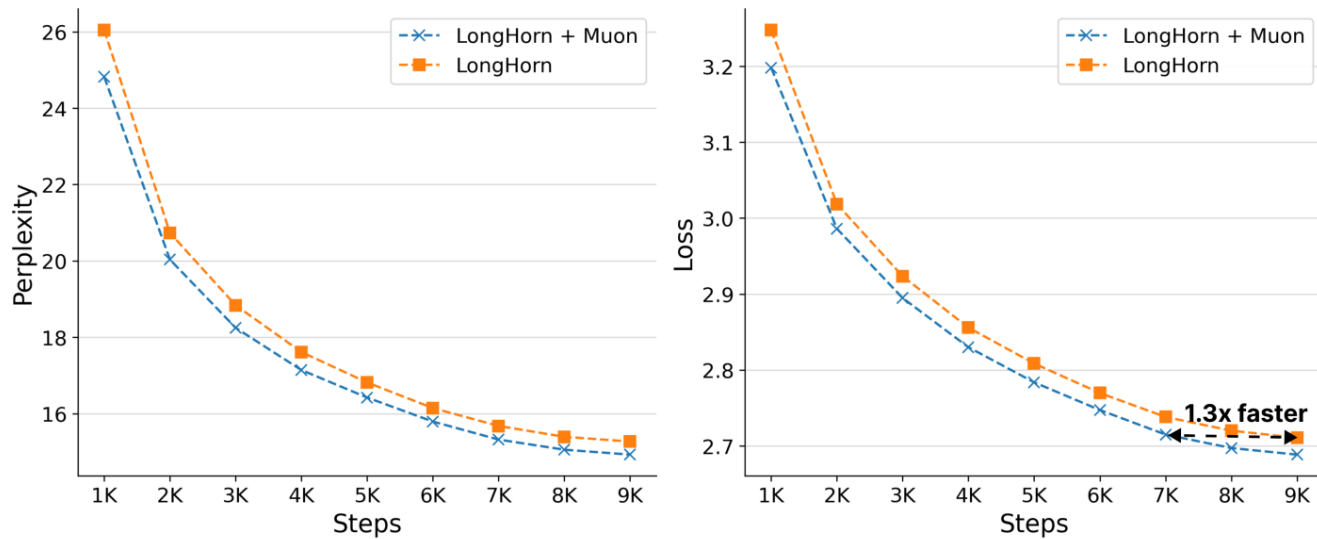
# Language Modeling and Long-Context Retrieval

Architecture	Memory Algorithm	S-NIAH-PK			S-NIAH-N			S-NIAH-UUID		
		2K	4K	8K	2K	4K	8K	2K	4K	8K
Mamba	Original	29.3	16.4	8.8	18.6	14.3	4.1	48.6	32.9	25.0
	+ Muon	<b>32.1</b>	<b>20.5</b>	<b>15.8</b>	<b>22.4</b>	<b>19.1</b>	<b>10.2</b>	<b>53.8</b>	<b>38.2</b>	<b>31.5</b>
LongHorn	Original	<b>67.9</b>	52.1	20.0	70.7	55.7	35.6	46.4	30.7	19.3
	+ Muon	66.7	<b>55.9</b>	<b>39.3</b>	<b>75.1</b>	<b>71.4</b>	<b>36.8</b>	<b>52.9</b>	<b>37.9</b>	28.6
GatedDeltaNet	Original	61.4	43.6	25.7	69.3	43.6	27.1	52.1	35.0	24.3
	+ Muon	<b>63.2</b>	<b>48.9</b>	<b>44.5</b>	<b>74.1</b>	<b>57.8</b>	<b>29.4</b>	<b>58.3</b>	<b>42.6</b>	<b>33.1</b>

**Table.** Long-context retrieval results on S-NIAH across 2K-8K context lengths.

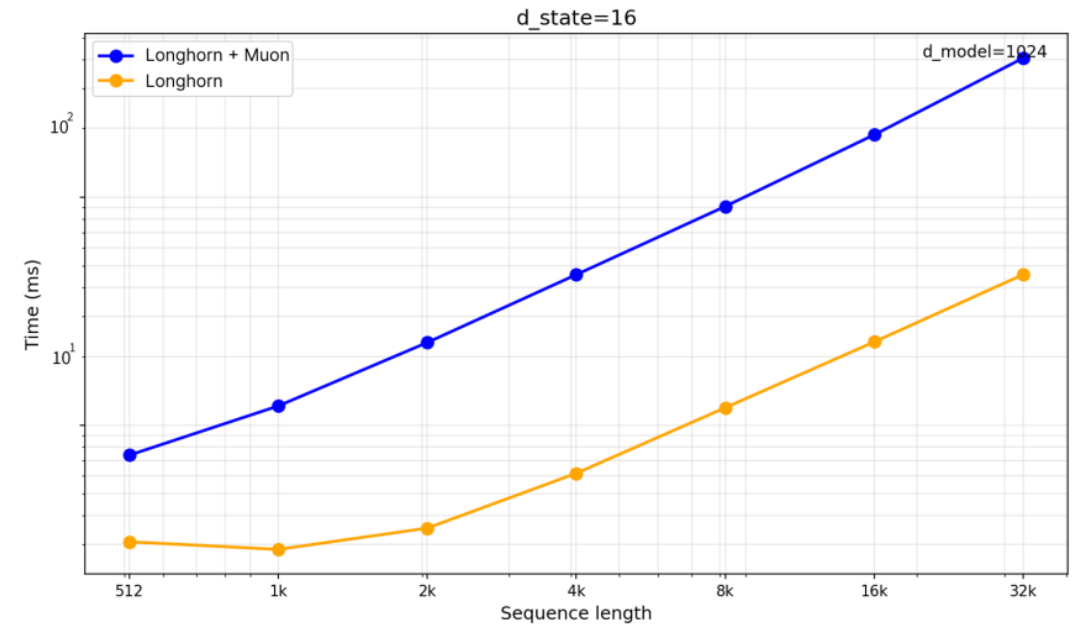
# Language Modeling and Long-Context Retrieval

*1.3x faster convergence*



**Figure a)** Pretraining dynamics on FineWeb-Edu 10B.

*Parallel Scan Preserved*



**Figure b)** Training time vs. sequence length.

# Vision Spatial Modeling and Robustness

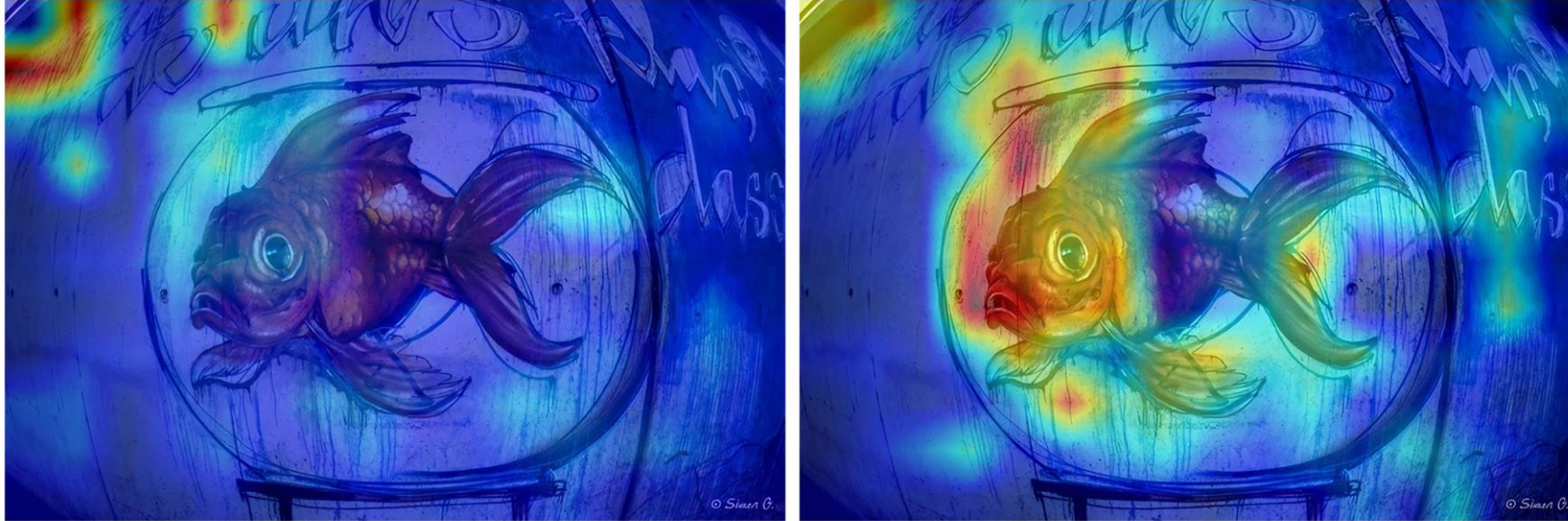
Architecture	Memory	IN-1K		IN-R	IN-A	IN-C	
	Algorithm	Top-1 $\uparrow$	Top-5 $\uparrow$	Top-1 $\uparrow$	Top-1 $\uparrow$	Top-1 $\uparrow$	mCE $\downarrow$
Mamba	Original	81.08	95.32	42.35	20.57	12.31	112.84
	<b>+ Muon</b>	<b>81.19</b>	<b>95.36</b>	<b>42.61</b>	20.50	<b>12.57</b>	<b>112.52</b>
LongHorn	Original	81.63	95.82	45.44	23.76	13.12	111.68
	<b>+ Muon</b>	<b>82.01</b>	<b>95.90</b>	<b>46.28</b>	<b>25.27</b>	<b>13.53</b>	<b>111.24</b>
GatedDeltaNet	Original	79.92	95.24	41.55	19.92	11.85	114.12
	<b>+ Muon</b>	<b>80.31</b>	<b>95.35</b>	<b>42.18</b>	<b>20.47</b>	<b>12.33</b>	<b>113.56</b>

**Table.** Image classification and robustness results on ImageNet.

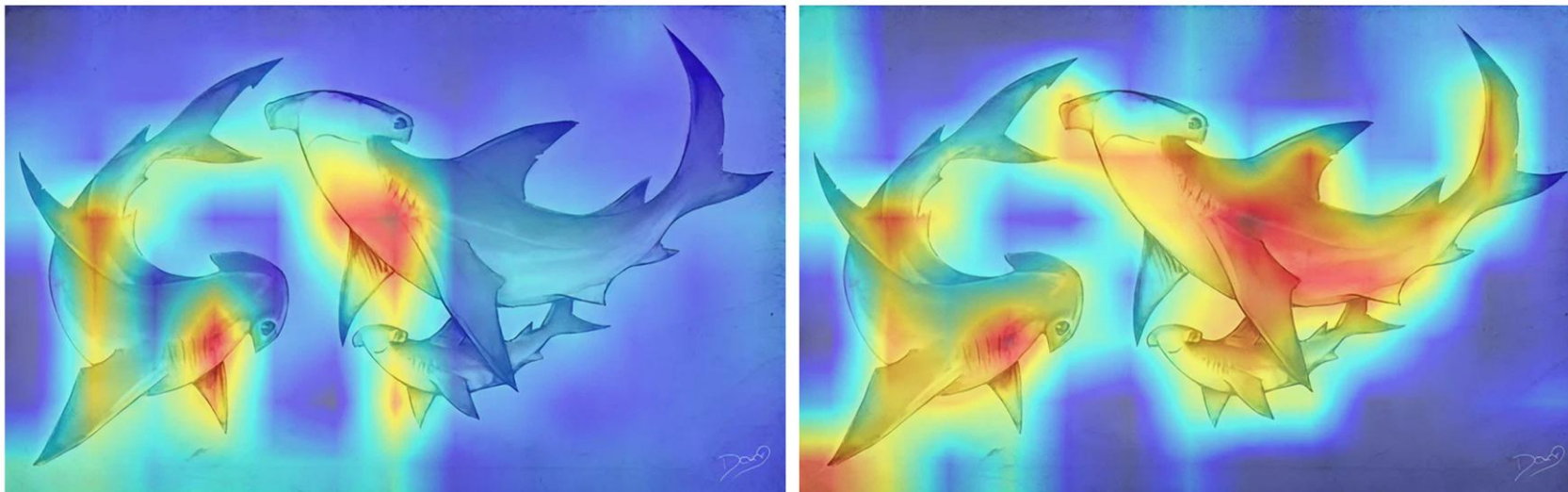
Architecture	Memory	Object Detection			Instance Seg.			Sem. Seg.
	Algorithm	AP <sup>box</sup>	AP <sub>50</sub> <sup>box</sup>	AP <sub>75</sub> <sup>box</sup>	AP <sup>mask</sup>	AP <sub>50</sub> <sup>mask</sup>	AP <sub>75</sub> <sup>mask</sup>	mIoU
Mamba	Original	50.8	69.5	55.2	44.1	67.0	47.9	43.9
	<b>+ Muon</b>	<b>51.1</b>	<b>69.9</b>	<b>55.4</b>	<b>44.3</b>	<b>67.4</b>	<b>48.2</b>	<b>45.2</b>
LongHorn	Original	50.6	69.3	55.3	44.0	66.7	47.6	44.2
	<b>+ Muon</b>	<b>51.0</b>	<b>69.8</b>	<b>55.4</b>	<b>44.1</b>	<b>67.1</b>	<b>47.6</b>	<b>45.7</b>
GatedDeltaNet	Original	49.5	68.1	53.8	43.4	67.2	46.8	41.2
	<b>+ Muon</b>	<b>50.1</b>	<b>68.8</b>	<b>54.5</b>	<b>43.8</b>	<b>67.8</b>	<b>47.3</b>	<b>41.8</b>

**Table.** Object detection and segmentation results on COCO and ADE20K.

# Vision Spatial Modeling and Robustness



**Figure.**  
*GradCAM  
visualizations*



(a) MambaVision

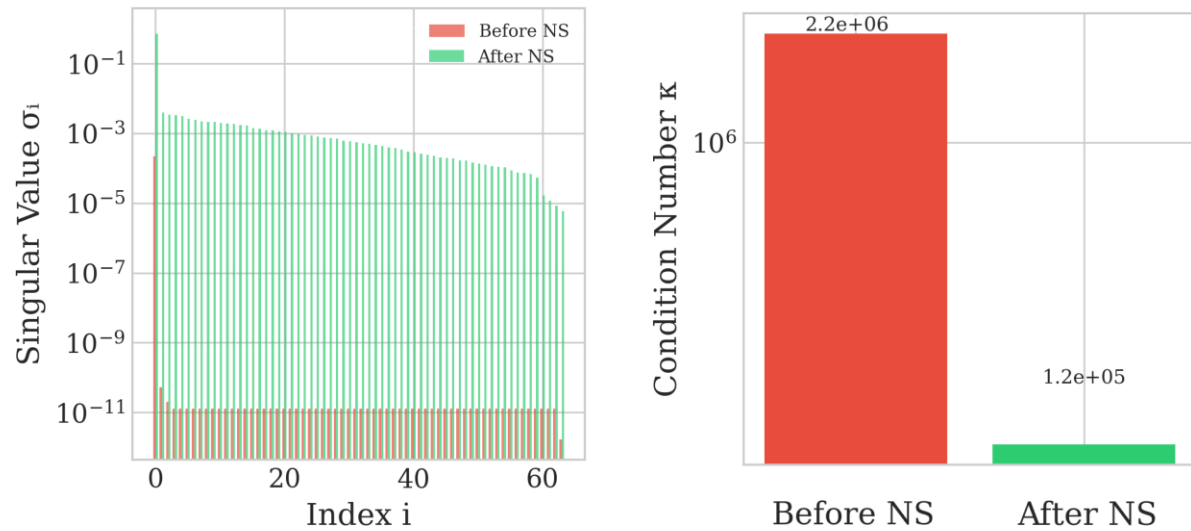
(b) MambaVision + Muon

# Time-Series for Human Activity Recognition

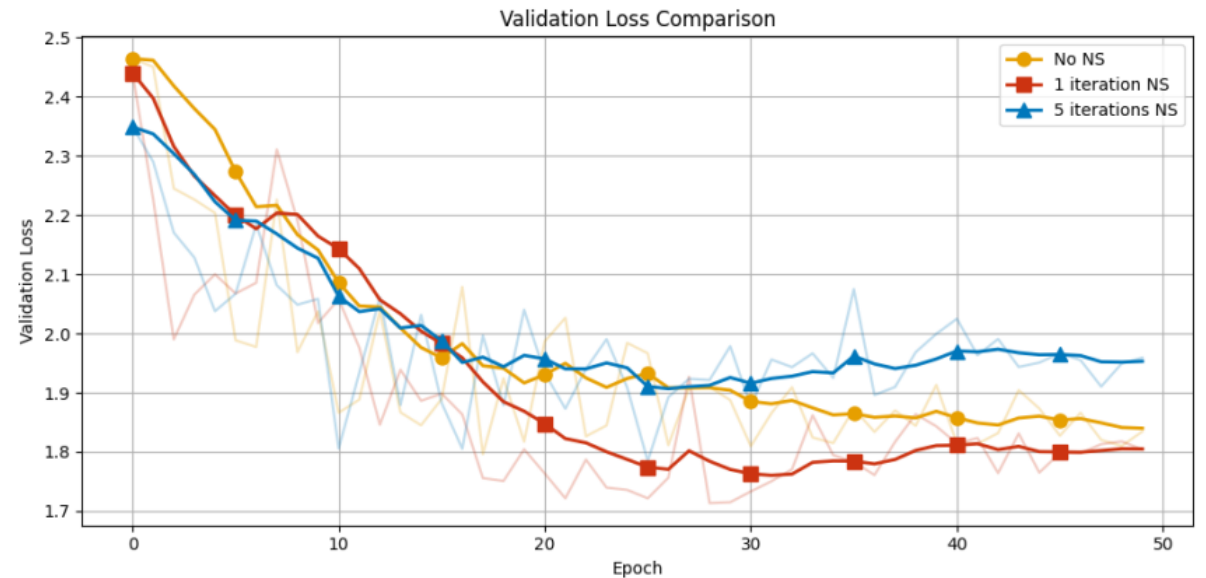
Architecture	Memory Algorithm	MuWiGes		UESTC-MMEA-CL		MMAct	
		Acc (%)	F1 (%)	Acc (%)	F1 (%)	Acc (%)	F1 (%)
Mamba	Original	96.25	96.25	87.74	88.91	71.46	71.68
	<b>+ Muon</b>	<b>97.64</b>	<b>97.44</b>	<b>91.62</b>	<b>90.96</b>	<b>74.65</b>	<b>74.25</b>
LongHorn	Original	97.23	97.35	89.06	89.43	72.47	73.76
	<b>+ Muon</b>	<b>97.95</b>	<b>97.96</b>	<b>91.56</b>	<b>91.02</b>	<b>74.40</b>	<b>76.43</b>
GatedDeltaNet	Original	96.88	96.87	86.09	85.92	66.39	67.75
	<b>+ Muon</b>	<b>97.73</b>	<b>97.73</b>	<b>87.97</b>	<b>87.81</b>	<b>66.61</b>	<b>68.73</b>

**Table.** Human activity recognition results on MuWiGes, UESTC-MMEA-CL, and MMAct.

# Impact of Newton-Schulz Iterations



**Figure a)** NS improves spectral conditioning by  $18\times$



**Figure b)** Ablation of NS iterations on MMAct dataset

# Capacity vs. Geometric Conditioning

*Do gains simply arise from increased capacity, since the augmented state  $\mathbf{Z}_t = [\mathbf{S}_t, \mathbf{M}_t]$  effectively doubles  $d_{state}$ ?*

Architecture	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
LongHorn	72.47 $\pm$ 0.19	75.68 $\pm$ 0.30	74.12 $\pm$ 0.17	73.76 $\pm$ 0.24
LongHorn $2 \times d_{state}$	72.88 $\pm$ 0.27	78.62 $\pm$ 0.34	74.67 $\pm$ 0.25	74.90 $\pm$ 0.23
<b>MuonLongHorn (Ours)</b>	<b>74.40 <math>\pm</math> 0.21</b>	<b>79.25 <math>\pm</math> 0.26</b>	<b>76.47 <math>\pm</math> 0.27</b>	<b>76.43 <math>\pm</math> 0.36</b>
Mamba	71.47 $\pm$ 0.25	71.82 $\pm$ 0.38	71.56 $\pm$ 0.28	71.68 $\pm$ 0.33
Mamba $2 \times d_{state}$	72.52 $\pm$ 0.22	76.98 $\pm$ 0.34	73.80 $\pm$ 0.25	73.40 $\pm$ 0.18
<b>MuonMamba (Ours)</b>	<b>74.65 <math>\pm</math> 0.34</b>	<b>78.16 <math>\pm</math> 0.29</b>	<b>74.06 <math>\pm</math> 0.41</b>	<b>74.25 <math>\pm</math> 0.35</b>

**Table.** MuonSSM outperforms doubled-state baselines on MMAAct.

# Conclusion

- **MuonSSM** stabilizes SSM training by conditioning the geometry of memory updates, without modifying the recurrent transition operator.
- A momentum pathway and Newton-Schulz normalization jointly improve gradient propagation, bound spectral growth, and enrich the effective rank of the memory state.
- The design preserves parallel scan compatibility –  $O(\log L)$  depth,  $O(L)$  work – with only a constant-factor overhead.
- Consistent gains across language, vision, and time-series establish geometric conditioning as a principled pathway to stable, scalable sequence modeling.
- **Future work:** larger-scale pretraining, hybrid attention-SSM architectures, adaptive conditioning strategies.

POSTER

Wed, Jul 8, 2026 • 10:30 AM – 12:15 PM KST

HALL A #1616



## MuonSSM: Orthogonalizing State Space Models for Sequence Modeling

Thai-Khanh Nguyen,<sup>\*1</sup> Ngoc-Bich-Uyen Vo,<sup>\*2</sup> Thieu N. Vo,<sup>3</sup> Tan M. Nguyen,<sup>4</sup> Cuong Pham,<sup>2</sup>  
<sup>1</sup>Dainam University, <sup>2</sup>Hanoi University of Science and Technology, <sup>3</sup>Posts and Telecommunications Institute of Technology, Vietnam  
<sup>4</sup>University of Bath, <sup>5</sup>National University of Singapore



### SSMs as Associative Memory

At each step, memory  $S_t \in \mathbb{R}^{d \times m}$  is updated from a key-value pair  $(k_t, v_t)$ :

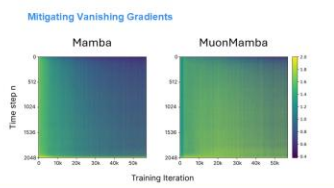
$$S_t = S_{t-1}(\alpha_t(I_m - \beta_t \eta k_t k_t^T)) + \beta_t v_t k_t^T$$

Model	$\alpha_t$	$\beta_t$	$\eta$	Update Rule
Mamba	1	0	1	$S_t = \alpha_t S_{t-1} + v_t k_t^T$
DeltaNet	1	$\beta_t$	1	$S_t = S_{t-1}(I - \beta_t k_t k_t^T) + \beta_t v_t k_t^T$
Gated DeltaNet	$\alpha_t$	$\beta_t$	1	$S_t = S_{t-1}(\alpha_t(I - \beta_t k_t k_t^T)) + \beta_t v_t k_t^T$
LongHor	1	$\frac{\beta_t}{1 + \beta_t k_t k_t^T}$	1	$S_t = S_{t-1}(I - \beta_t k_t k_t^T) + \beta_t v_t k_t^T$

### First-Order Updates Degrade

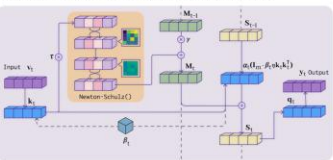
Rank-one writes are restricted to the current key direction. Over long horizons, this causes:

- Gradient degradation:** Vanishing gradients through  $\prod_{n=t}^T \alpha_n (I_m - \beta_n \eta k_n k_n^T)$ .
- Spectral anisotropy:** Singular values become highly non-uniform.
- Memory interference:** New updates can overwrite previous information.



### MuonSSM

Two lightweight conditioning components:



1. Single-iteration Newton-Schulz (NS): spectral conditioning

$$\tilde{X} = \frac{X}{\max(\|X\|_F, \delta)}$$

$$NS(X) = (a + b\tilde{X}\tilde{X}^T + c(\tilde{X}\tilde{X}^T)^2)\tilde{X}$$

$(a, b, c) = (3.4445, -4.7750, 2.0315), \delta > 0$

2. Momentum ( $M_t$ ): temporal accumulation

$$M_t = \gamma M_{t-1} + NS(\tau \beta_t v_t k_t^T)$$

$$S_t = S_{t-1}(\alpha_t(I_m - \beta_t \eta k_t k_t^T)) + M_t$$

$\tau > 0; \gamma \in (0, 1]$

### Key Takeaways

- Stable long-range memory via geometric conditioning.
- A unified, backbone-agnostic framework for modern SSMs.

### Parallelizability of MuonSSM

The coupled  $Z_t = [S_t, M_t] \in \mathbb{R}^{d \times 2m}$  follows block-affine recurrence  $Z_t = Z_{t-1}\Phi_t + \Psi_t$ , enabling parallel scans with  $O(\log L)$  depth,  $O(L)$  work.

$$\Phi_t = \begin{bmatrix} \alpha_t(I_m - \beta_t \eta k_t k_t^T) & 0 \\ \gamma I & \gamma I \end{bmatrix}; \Psi_t = \begin{bmatrix} NS(\tau \beta_t v_t k_t^T) \\ NS(\tau \beta_t v_t k_t^T) \end{bmatrix}$$

### Theoretical Analysis

#### Gradient Stability

The gradient of the loss  $\mathcal{L}$  with respect to  $Z_{t-1}$

$$\frac{\partial \mathcal{L}}{\partial Z_{t-1}} = \frac{\partial \mathcal{L}}{\partial Z_t} \left[ \prod_{n=t}^T D_n^T \sum_{k=t}^T \left( \prod_{j=t}^{k+1} D_j^T \right) (y_{1m})^{k-t+1} \right]$$

When  $\gamma \approx 1$ , momentum pathway  $(\gamma I_m)^{T-t+1}$  preserves long-range gradients.

#### Rank Enrichment

- NS biases writes off the rank-1 direction
- Momentum accumulates the NS writes
- higher effective rank, richer memory.

### Conclusion

- MuonSSM conditions memory updates via momentum and spectral normalization.
- MuonSSM improves gradients and state rank while preserving parallel scans.
- Future work: scaling, hybrid models, and adaptive conditioning.

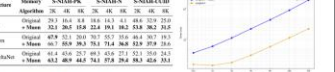
### Empirical Results

#### Language Modeling and Long-Context Retrieval

Zero-shot performance – FineWeb-Edu10B tokens

Architecture	Memory	NS	NS+M	NS+M+P	NS+M+P+R	NS+M+P+R+L	NS+M+P+R+L+G	NS+M+P+R+L+G+D	NS+M+P+R+L+G+D+V	NS+M+P+R+L+G+D+V+T	NS+M+P+R+L+G+D+V+T+K
Mamba	Original	22.17	22.22	22.27	22.32	22.37	22.42	22.47	22.52	22.57	22.62
	+ Muon	22.85	22.90	22.95	23.00	23.05	23.10	23.15	23.20	23.25	23.30
LongHor	Original	21.80	21.85	21.90	21.95	22.00	22.05	22.10	22.15	22.20	22.25
	+ Muon	22.50	22.55	22.60	22.65	22.70	22.75	22.80	22.85	22.90	22.95
Gated DeltaNet	Original	20.50	20.55	20.60	20.65	20.70	20.75	20.80	20.85	20.90	20.95
	+ Muon	21.20	21.25	21.30	21.35	21.40	21.45	21.50	21.55	21.60	21.65

#### Needle-in-a-Haystack



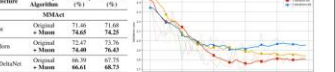
#### Vision and Robustness

##### Image Classification on ImageNet

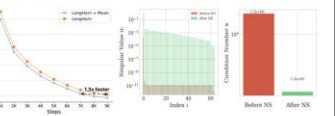
Architecture	Memory	NS	NS+M	NS+M+P	NS+M+P+R	NS+M+P+R+L	NS+M+P+R+L+G	NS+M+P+R+L+G+D	NS+M+P+R+L+G+D+V	NS+M+P+R+L+G+D+V+T	NS+M+P+R+L+G+D+V+T+K
Mamba	Original	81.00	81.05	81.10	81.15	81.20	81.25	81.30	81.35	81.40	81.45
	+ Muon	81.40	81.45	81.50	81.55	81.60	81.65	81.70	81.75	81.80	81.85
LongHor	Original	81.40	81.45	81.50	81.55	81.60	81.65	81.70	81.75	81.80	81.85
	+ Muon	81.80	81.85	81.90	81.95	82.00	82.05	82.10	82.15	82.20	82.25
GatedDeltaNet	Original	80.50	80.55	80.60	80.65	80.70	80.75	80.80	80.85	80.90	80.95
	+ Muon	81.00	81.05	81.10	81.15	81.20	81.25	81.30	81.35	81.40	81.45

#### Time-Series

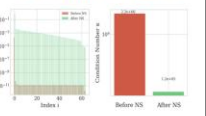
##### Human Activity Recognition



#### 1.3x faster convergence



#### Improved Singular-Value Spectrum



Poster



Paper

### Our labs

<https://tanmnguyen89.github.io>

<https://sites.google.com/view/cuongpham>

**Reference:** Nguyen, Thai-Khanh, Vo, Ngoc-Bich-Uyen, Vo, Thieu N., Nguyen, Tan M., and Pham, Cuong. MuonSSM: Orthogonalizing State Space Models for Sequence Modeling, *International Conference on Machine Learning*, 2026. <https://github.com/t-khanusa/MuonSSM>

**THANKS FOR  
YOUR ATTENTION!**