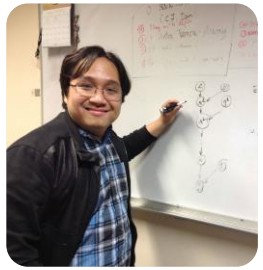
ICML
International Conference
On Machine Learning

MuonSSM:
Orthogonalizing State Space Models for Sequence Modeling

Thai-Khanh Nguyen[1,3*] **Ngoc-Bich-Uyen Vo[2*]** Thieu N. Vo[4†] Tan M. Nguyen[5†] Cuong Pham[2†]




[1] Dainam University
[2] Posts and Telecommunications Institute of Technology
[3] Hanoi University of Science and Technology
[4] University of Bath
[5] 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[1]:

Updated memory state:

$$\underbrace{\mathbf{S}_t}_{\mathbb{R}^{d\times m}} = \mathbf{S}_{t-1}\left(\alpha_t\left(\mathbf{I}_m - \beta_t \eta \mathbf{k}_t \mathbf{k}_t^\top\right)\right) + \beta_t \underbrace{\mathbf{v}_t}_{\mathbb{R}^d}\underbrace{\mathbf{k}_t^\top}_{\mathbb{R}^m}, \qquad (1)$$

recall correction

$(0, 1]$ memory retention

> 0 update magnitude

(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:

$$\underbrace{\mathbf{S}_t}_{\mathbb{R}^{d\times m}} = \mathbf{S}_{t-1}\Big(\underbrace{\alpha_t}_{(0,1]\text{ memory retention}}\big(\mathbf{I}_m - \underbrace{\beta_t}_{>0\text{ update magnitude}}\eta\,\mathbf{k}_t\mathbf{k}_t^\top\big)\Big) + \beta_t\underbrace{\mathbf{v}_t}_{\mathbb{R}^d}\underbrace{\mathbf{k}_t^\top}_{\mathbb{R}^m}, \qquad (1)$$

recall correction

Model	α_t	β_t	η	Update Rule
Mamba (Gu & Dao, 2024)	α_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	β_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)	α_t	β_t	1	$\mathbf{S}_t = \mathbf{S}_{t-1}\big(\alpha_t(\mathbf{I} - \beta_t\mathbf{k}_t\mathbf{k}_t^\top)\big) + \beta_t\mathbf{v}_t\mathbf{k}_t^\top$
LongHorn (Liu et al., 2024)	1	$\frac{\beta_t}{1+\beta_t\mathbf{k}_t^\top\mathbf{k}_t}$	1	$\mathbf{S}_t = \mathbf{S}_{t-1}\big(\mathbf{I} - \frac{\beta_t}{1+\beta_t\mathbf{k}_t^\top\mathbf{k}_t}\mathbf{k}_t\mathbf{k}_t^\top\big) + \frac{\beta_t}{1+\beta_t\mathbf{k}_t^\top\mathbf{k}_t}\mathbf{v}_t\mathbf{k}_t^\top$

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.

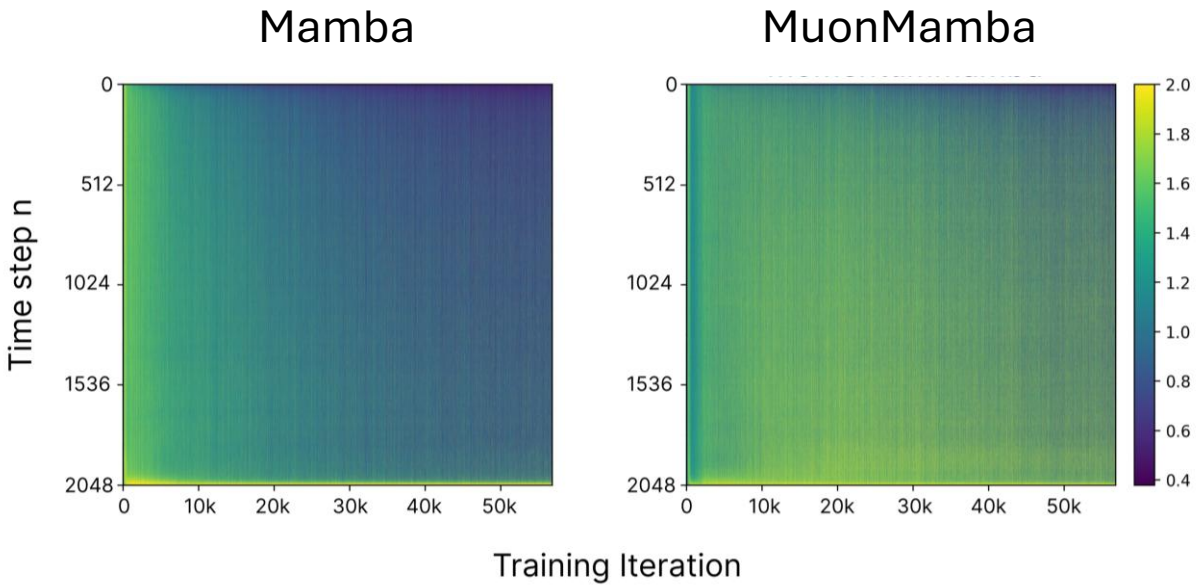


Figure 1. *MuonMamba (our method) shows more uniform gradient flow over long contexts than Mamba.*

MuonSSM: Orthogonalizing State Space Models

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

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

MuonSSM

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

Single-iteration Newton-Schulz

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

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

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

$$\mathbf{NS}(\boldsymbol{X}) = \left(a + b\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^{\boldsymbol{T}} + c\left(\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^{\boldsymbol{T}}\right)^2\right)\widetilde{\boldsymbol{X}}$$

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

***MuonOptimizer*[2]**

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

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

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

MuonSSM

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

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

MuonSSM: Orthogonalizing State Space Models

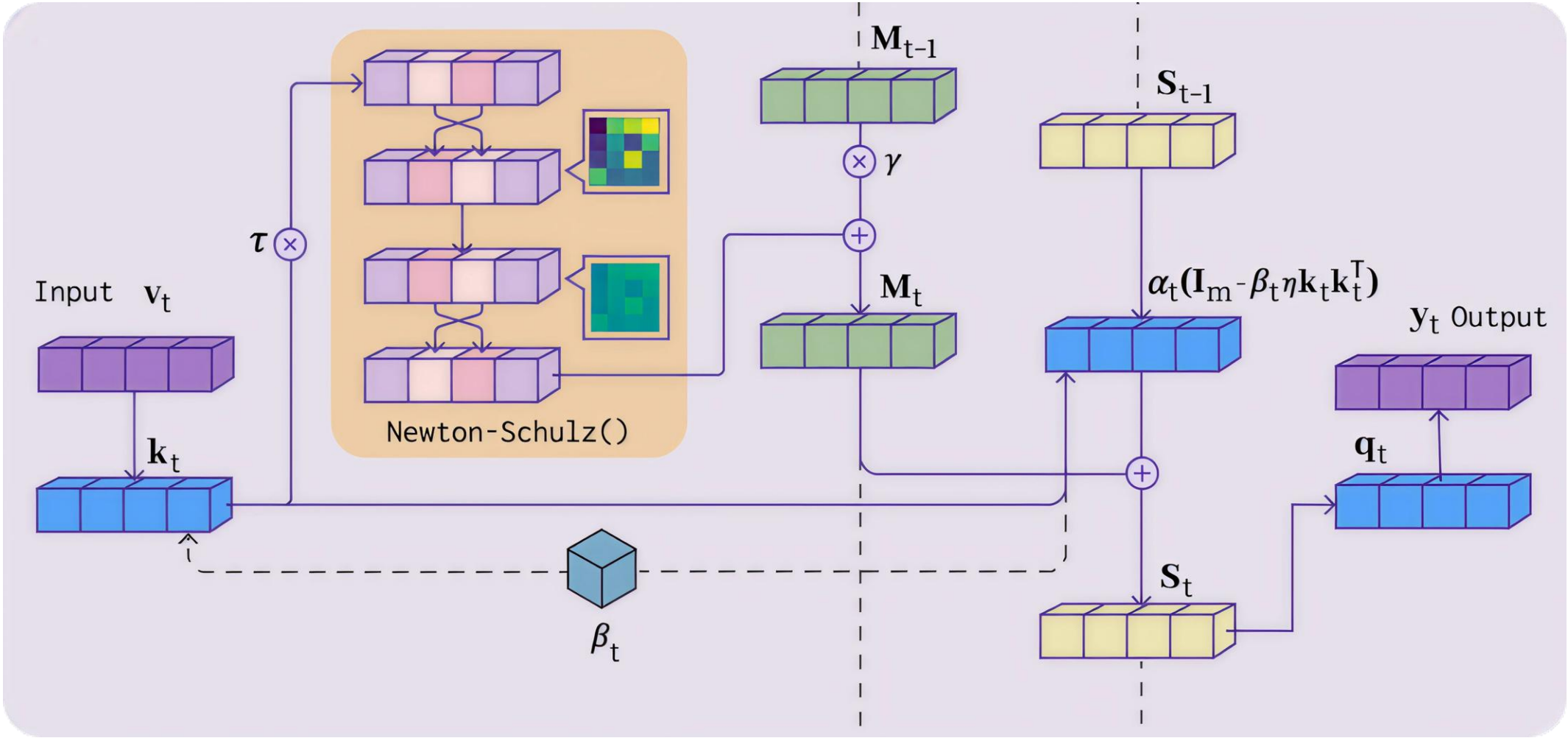


Figure. *The MuonSSM architecture.*

Parallelizability of MuonSSM

Block-Affine Recurrence

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

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

$$\mathbf{Z}_t = [\mathbf{S}_t \quad \mathbf{M}_t] \in \mathbb{R}^{d \times 2m}; \boldsymbol{\Phi}_t = \begin{bmatrix} \alpha_t\left(\mathbf{I_m} - \beta_t \eta \boldsymbol{k}_t \boldsymbol{k}_t^T\right) & \mathbf{0} \\ \gamma \mathbf{I} & \gamma \mathbf{I} \end{bmatrix}; \boldsymbol{\Psi}_t = \begin{bmatrix} \mathbf{NS}\left(\tau \beta_t \boldsymbol{v}_t \boldsymbol{k}_t^T\right) \\ \mathbf{NS}\left(\tau \beta_t \boldsymbol{v}_t \boldsymbol{k}_t^T\right) \end{bmatrix}$$

$$\mathbf{Z}_t = \mathbf{Z}_{t-1} \boldsymbol{\Phi}_t + \boldsymbol{\Psi}_t$$

(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

$$\mathbf{Z}_t = \mathbf{Z}_{t-1}\mathbf{\Phi}_t + \mathbf{\Psi}_t$$

$$\mathbf{\Phi}_t = \begin{bmatrix} \mathbf{D}_t & \mathbf{0} \\ \gamma\mathbf{I} & \gamma\mathbf{I} \end{bmatrix}; \quad \mathbf{D}_t = \alpha_t\left(\mathbf{I_m} - \beta_t\eta\boldsymbol{k}_t\boldsymbol{k}_t^{\boldsymbol{T}}\right)$$

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

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**

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

***Backward* Geometry of Newton-Schulz Normalization**

For rank-one $\mathbf{X_0} = uw^{\mathrm{T}}$, the NS Jacobian decomposes into four orthogonal eigenspaces.

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

- **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, MMAct

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)	**40.83**	**89.17**	**22.84**	63.47	**33.19**	**53.36**	**53.21**	**25.58**	**38.33**	**63.82**	**44.23**
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)	**41.71**	**80.98**	**24.00**	62.02	**32.85**	**54.38**	**51.17**	**25.12**	**37.15**	**59.44**	**43.27**
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)	**38.12**	**83.47**	**23.91**	**62.88**	**33.51**	**53.76**	**52.74**	**23.13**	**38.02**	**56.85**	**43.10**

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

Language Modeling and Long-Context Retrieval

Architecture	Memory	S-NIAH-PK			S-NIAH-N			S-NIAH-UUID		
	Algorithm	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	**32.1**	**20.5**	**15.8**	**22.4**	**19.1**	**10.2**	**53.8**	**38.2**	**31.5**
LongHorn	Original	**67.9**	52.1	20.0	70.7	55.7	35.6	46.4	30.7	19.3
	+ Muon	66.7	**55.9**	**39.3**	**75.1**	**71.4**	**36.8**	**52.9**	**37.9**	28.6
GatedDeltaNet	Original	61.4	43.6	25.7	69.3	43.6	27.1	52.1	35.0	24.3
	+ Muon	**63.2**	**48.9**	**44.5**	**74.1**	**57.8**	**29.4**	**58.3**	**42.6**	**33.1**

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

Language Modeling and Long-Context Retrieval

1.3x faster convergence

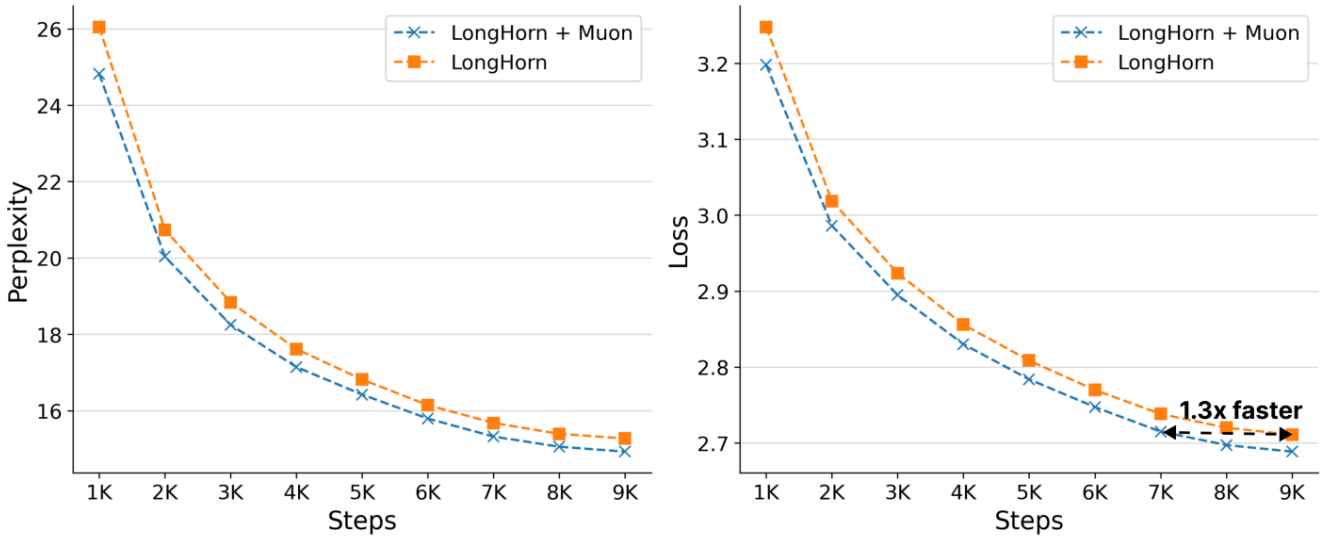


Figure a) *Pretraining dynamics on FineWeb-Edu 10B.*

Parallel Scan Preserved

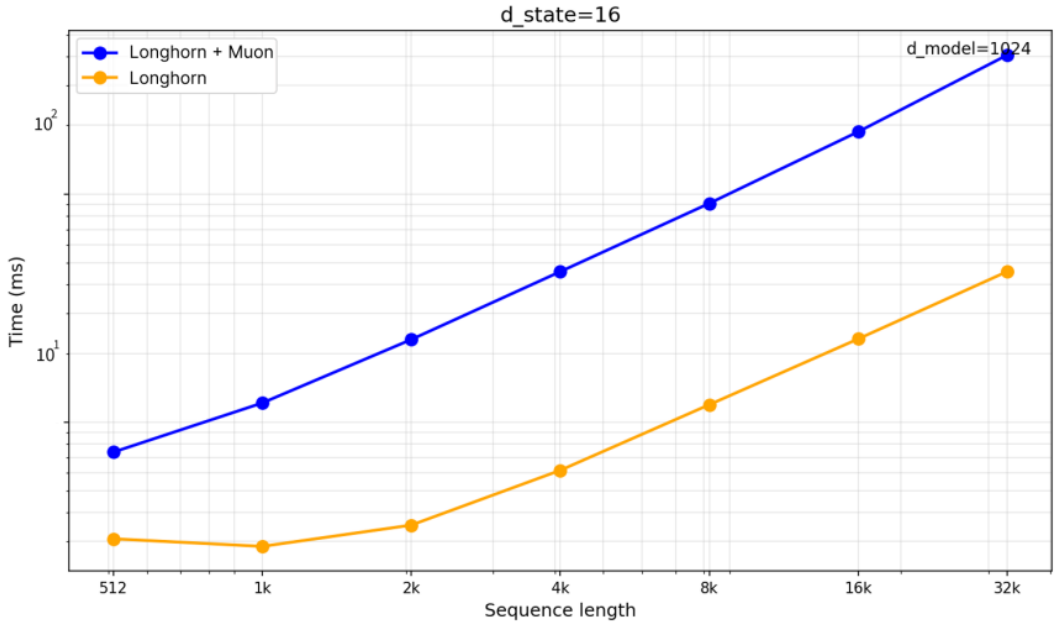


Figure b) *Training time vs. sequence length.*

Vision Spatial Modeling and Robustness

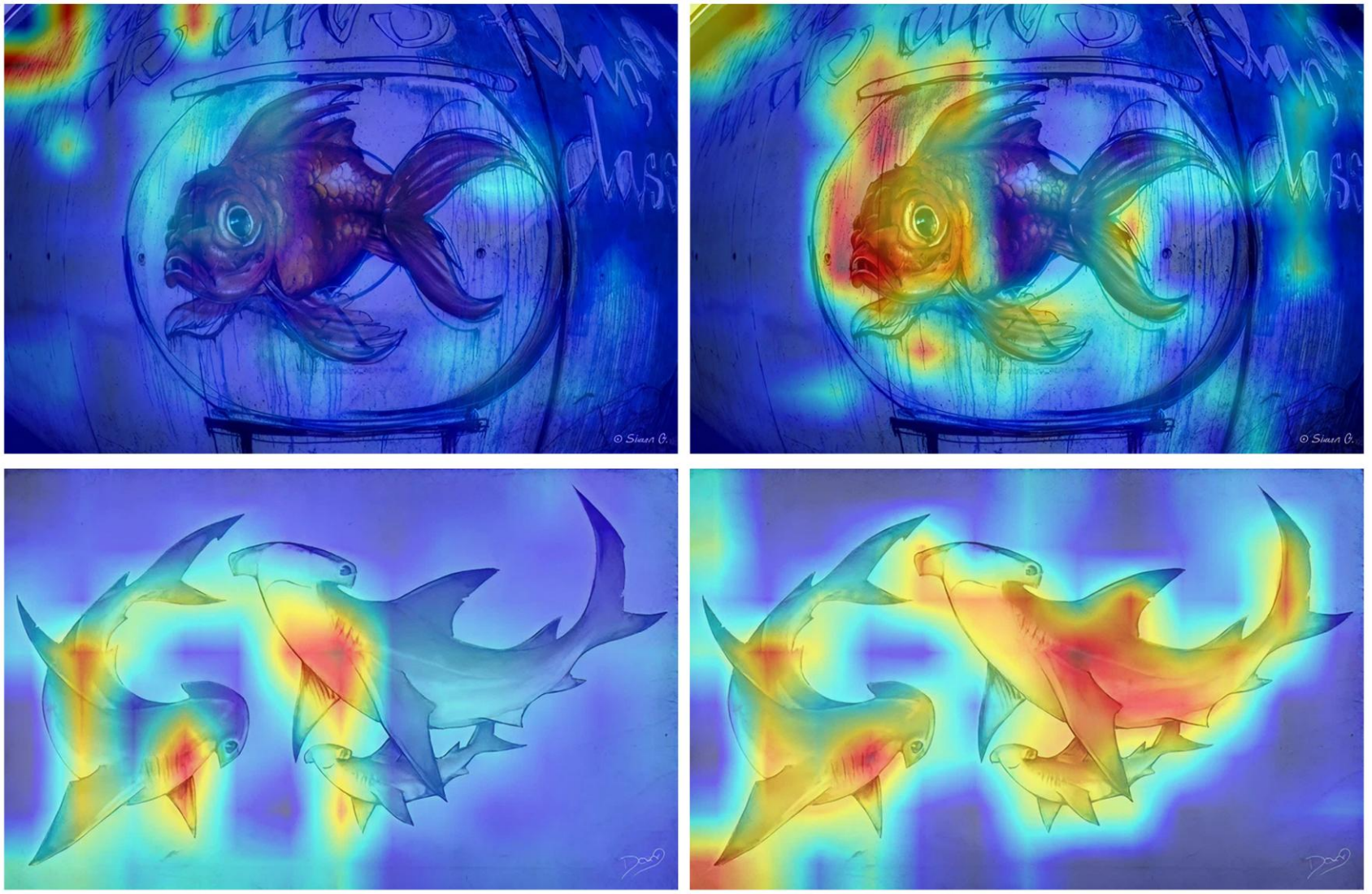
Architecture	Memory	IN-1K		IN-R	IN-A	IN-C	
	Algorithm	Top-1 ↑	Top-5 ↑	Top-1 ↑	Top-1 ↑	Top-1 ↑	mCE ↓
Mamba	Original	81.08	95.32	42.35	20.57	12.31	112.84
	+ Muon	**81.19**	**95.36**	**42.61**	20.50	**12.57**	**112.52**
LongHorn	Original	81.63	95.82	45.44	23.76	13.12	111.68
	+ Muon	**82.01**	**95.90**	**46.28**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

Table. Image classification and robustness results on ImageNet.

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

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

Vision Spatial Modeling and Robustness



(a) MambaVision (b) MambaVision + Muon

Figure. *GradCAM visualizations*

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
	+ Muon	**97.64**	**97.44**	**91.62**	**90.96**	**74.65**	**74.25**
LongHorn	Original	97.23	97.35	89.06	89.43	72.47	73.76
	+ Muon	**97.95**	**97.96**	**91.56**	**91.02**	**74.40**	**76.43**
GatedDeltaNet	Original	96.88	96.87	86.09	85.92	66.39	67.75
	+ Muon	**97.73**	**97.73**	**87.97**	**87.81**	**66.61**	**68.73**

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

Impact of Newton-Schulz Iterations

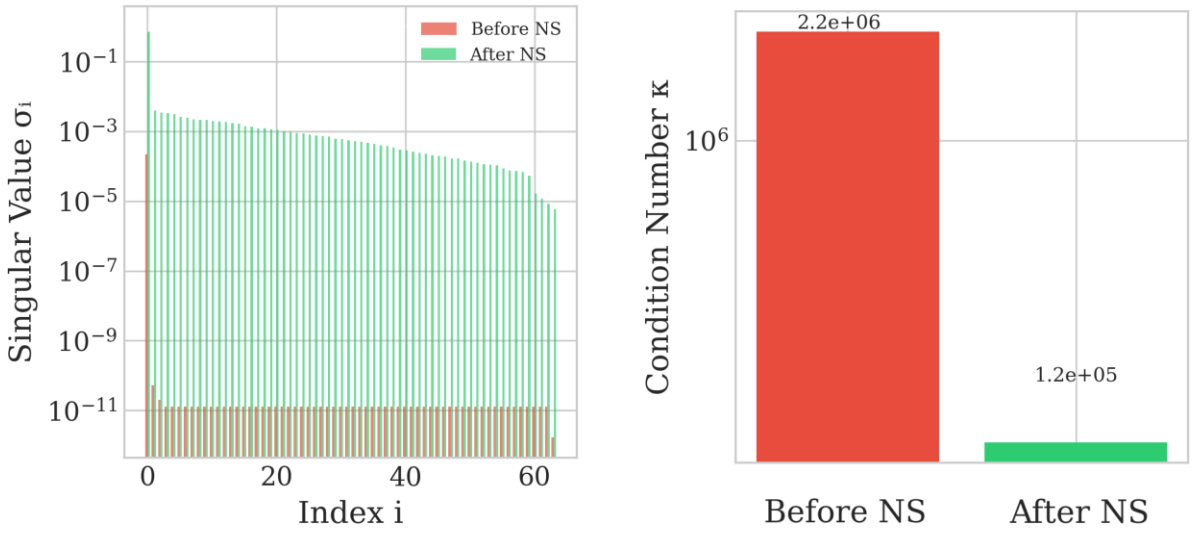


Figure a) *NS improves spectral conditioning by 18×*

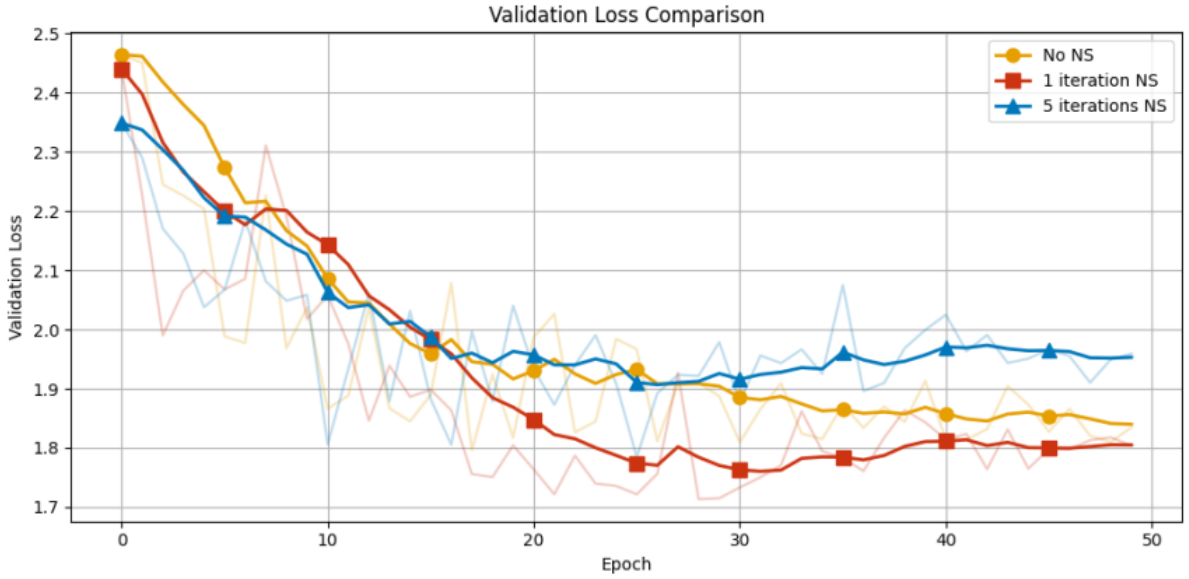


Figure b) *Ablation of NS iterations on MMAct dataset*

Capacity vs. Geometric Conditioning

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

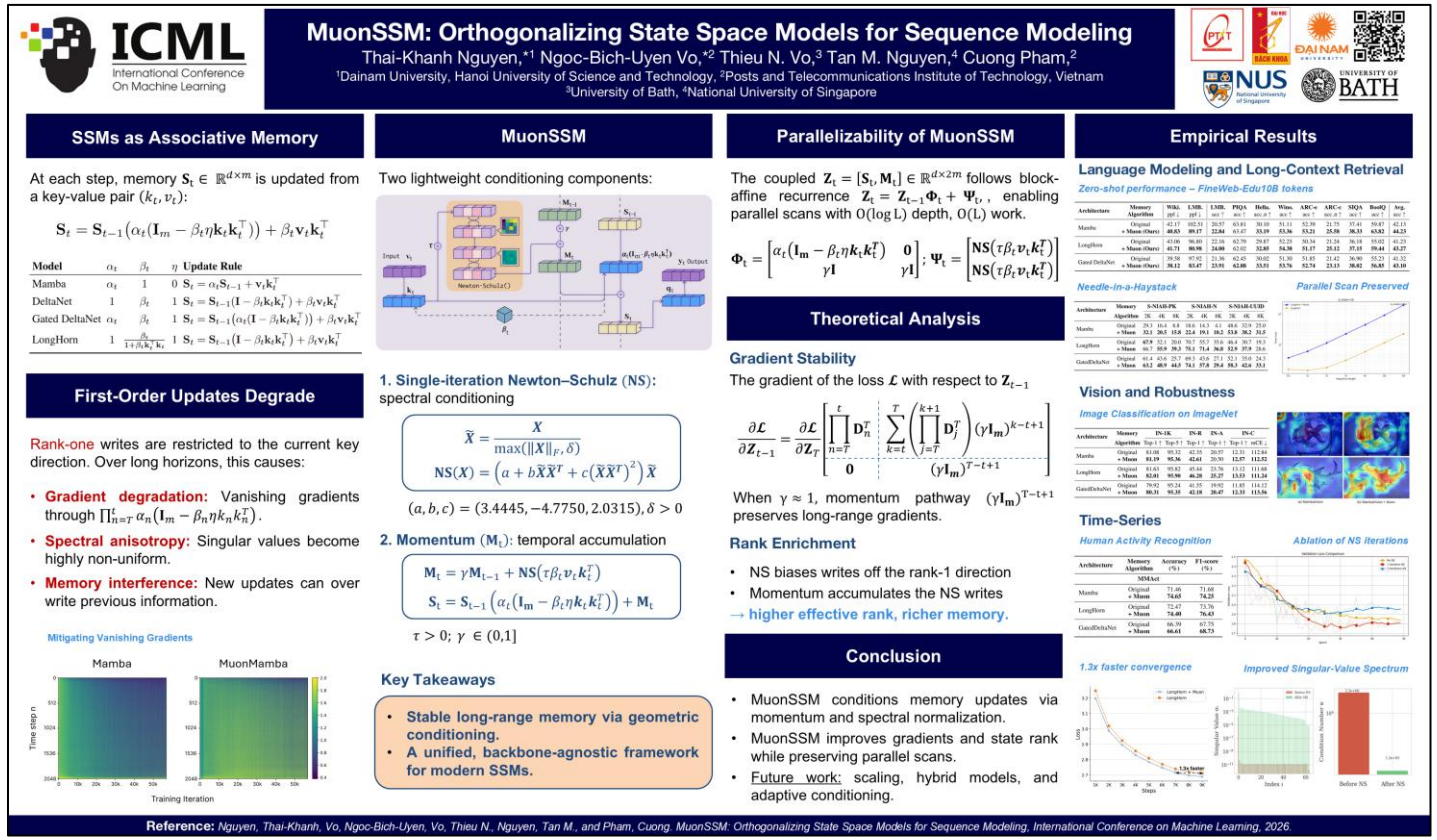
Architecture	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
LongHorn	72.47 ± 0.19	75.68 ± 0.30	74.12 ± 0.17	73.76 ± 0.24
LongHorn $2\times d_{\text{state}}$	72.88 ± 0.27	78.62 ± 0.34	74.67 ± 0.25	74.90 ± 0.23
MuonLongHorn (Ours)	$\mathbf{74.40 \pm 0.21}$	$\mathbf{79.25 \pm 0.26}$	$\mathbf{76.47 \pm 0.27}$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

Table. MuonSSM outperforms doubled-state baselines on MMAct.

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



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!