A Theoretical Analysis of Mamba's Training Dynamics: Filtering Relevant Features for Generalization in State Space Models

Mugunthan Shandirasegaran[1], Hongkang Li[2], Songyang Zhang[3], Meng Wang[4], Shuai Zhang[1]

[1]New Jersey Institute of Technology

[2]University of Pennsylvania

[3]University of Louisiana at Lafayette

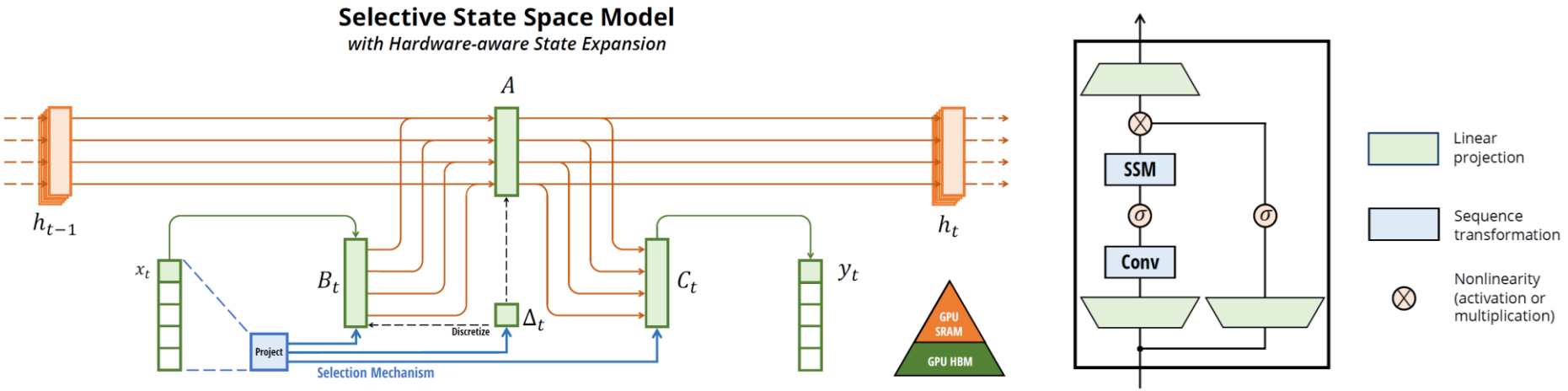
[4]Rensselaer Polytechnic Institute

International Conference on Learning Representations (ICLR 2026)
April 2026

ICLR

Mamba

- Mamba is a recent State Space Model (SSM) which introduces **selective gating**.
- It renewed interest because of linear computational complexity and demonstrated strong empirical performance.

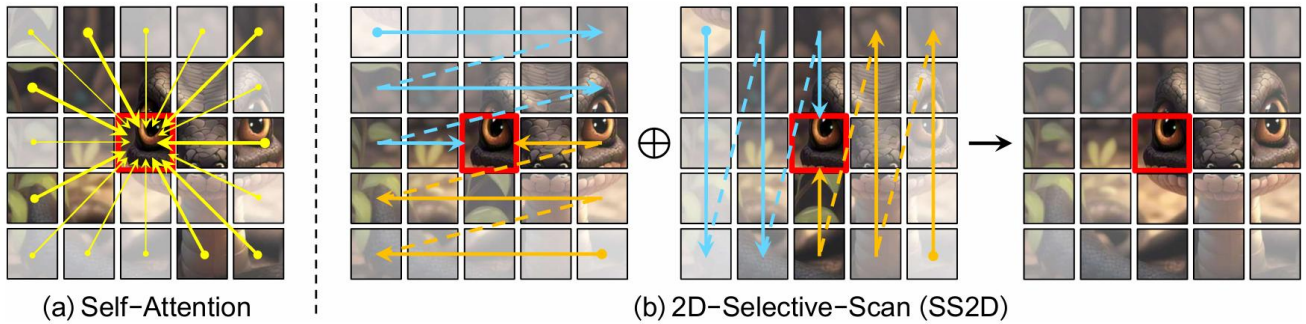


Source: Gu & Dao, 2024

Lack of Theoretical Understanding

- Visual State Space Model highlighted the **role of token order and scanning strategies.**

VMamba: Visual State Space Model (NeurIPS 2024)



Source: Liu et al., 2024

- Despite the empirical success, the theoretical understanding remains less investigated.
  - What types of data distributions can Mamba learn?
  - How does Mamba's gating affect representation learning compared to Transformers?

Highlights of Our Theoretical Insights

- Characterization of the optimization trajectories of the gating weights
  - We prove that the gating mechanism selects class-relevant features while ignoring irrelevant ones.

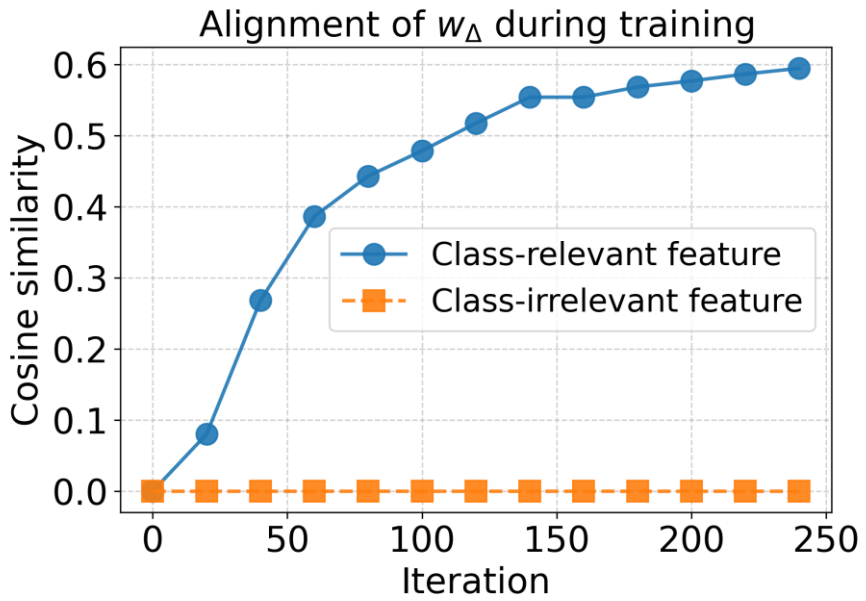
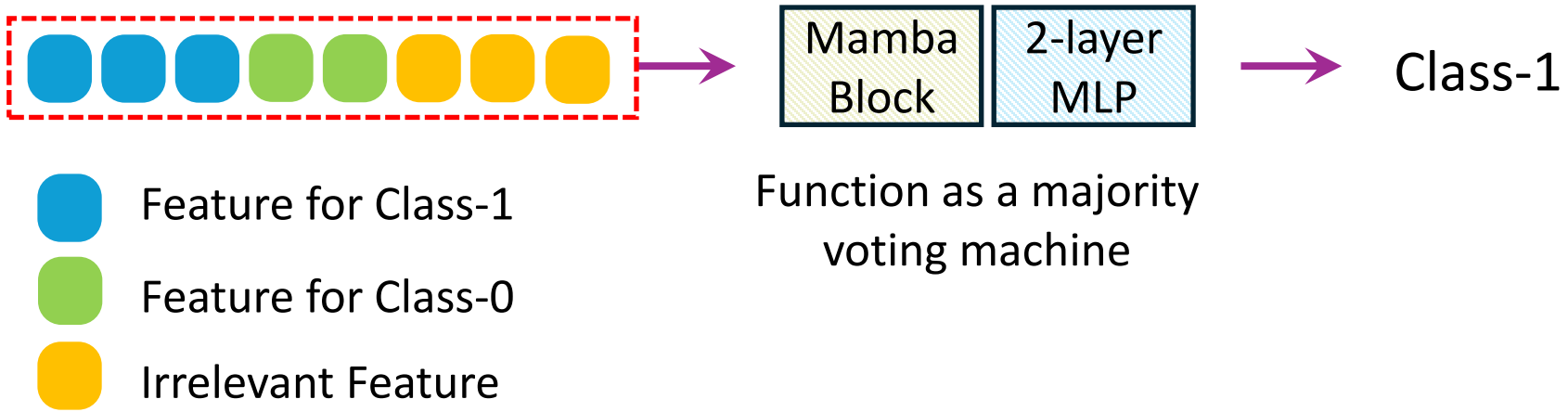


Fig 1: Alignment of $\boldsymbol{w_\Delta}$ during training

Highlights of Our Theoretical Insights

- Characterization of the optimization trajectories of the gating weights
  - We prove that the gating mechanism selects class-relevant features while ignoring irrelevant ones.
- Mamba can learn from data where the label is determined by a **majority vote over class-relevant features.**



Highlights of Our Theoretical Insights

- Mamba can learn locality-structured data, where class-relevant features appear in spatially or temporally concentrated regions.
  - Higher local concentration of class-relevant features accelerates learning.

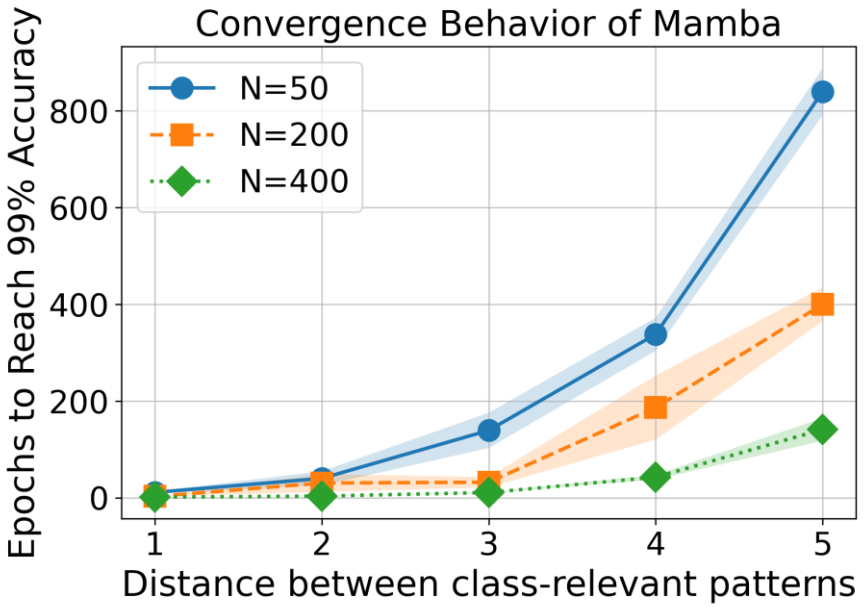


Fig 2: Convergence for **Locality-structured data**

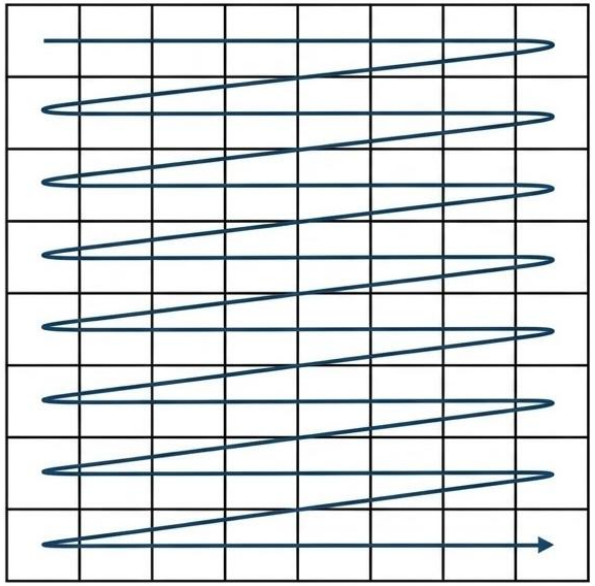
Which Scan is better?

- Our results help explain why scan orders that preserve local structure, such as Hilbert curves, lead to improved performance.

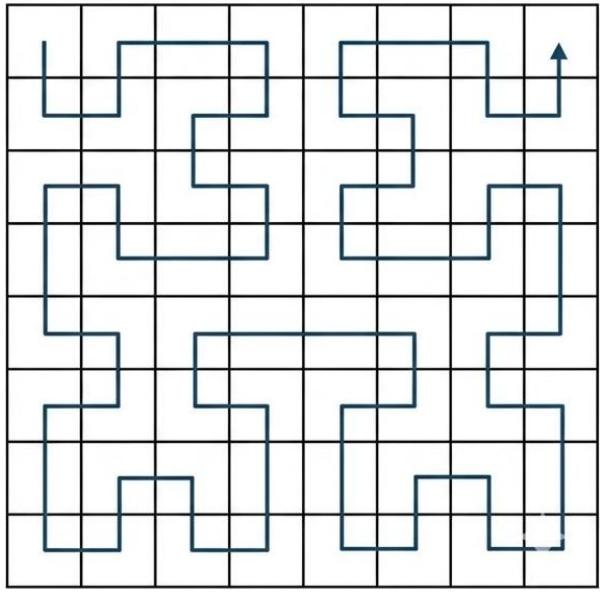
Order	Accuracy
Horizontal	0.870 ± 0.024
Hilbert	0.893 ± 0.019

Results on whole slide image classification from (Jiang et al. 2026).

Horizontal (row-wise) Scan



Hilbert Curve Scan



Performance on Locality-Structured Data

- Mamba outperforms Transformer and local attention.
  - Global attention performs near random.
  - Local attention learns meaningful patterns but remains less effective than Mamba.

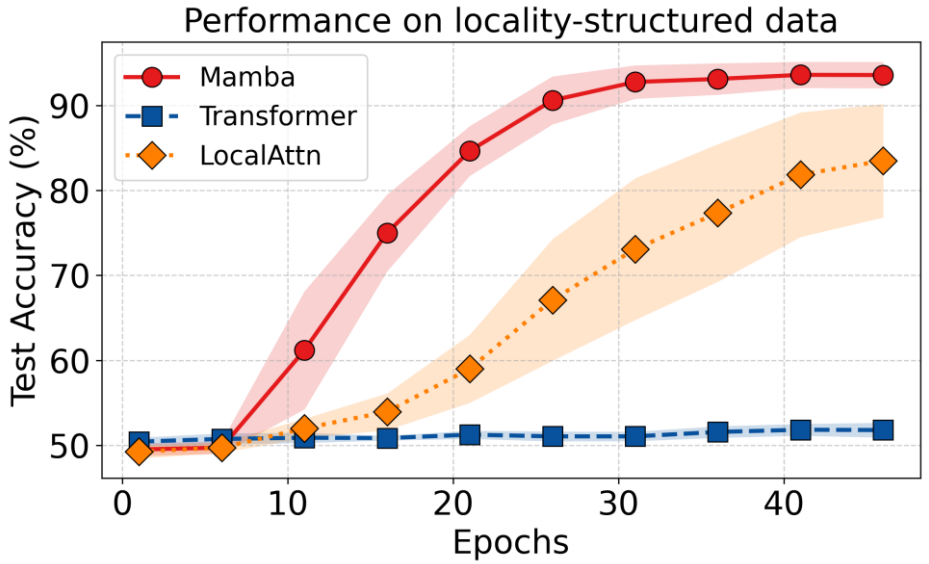


Fig 3: Mamba outperforms on **Locality-structured data**