2DMamba: Efficient State Space Model for Image Representation with Applications on Giga-Pixel Whole Slide Image Classification

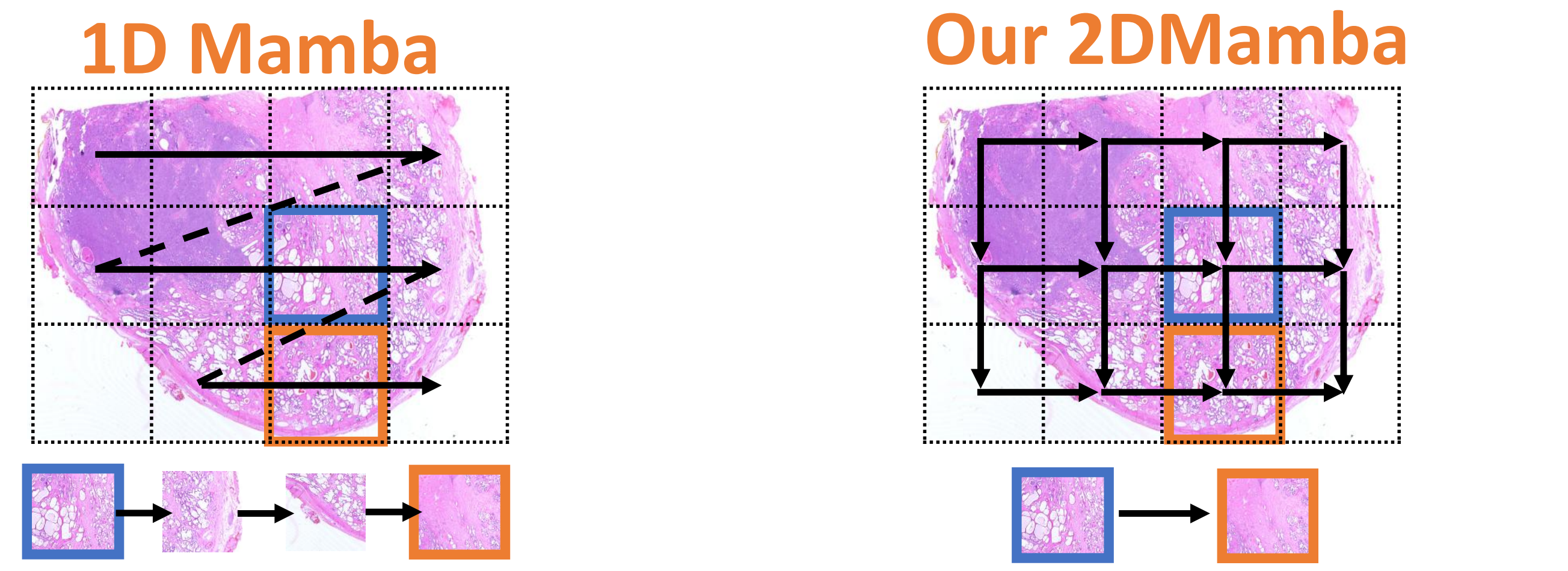
Jingwei Zhang[1]*, Anh Tien Nguyen[2]*, Xi Han[1]*, Vincent Quoc-Huy Trinh[5], Hong Qin[1], Dimitris Samaras[1], Mahdi S. Hosseini[3,4]

[1]Stony Brook University; [2]Korea University; [3]Concordia University; [4]Mila–Quebec AI Institute; [5]University of Montreal Hospital Center

CVPR Nashville JUNE 11-15, 2025

Introduction

- Mamba [1], a state space model (SSM) with **linear time** complexity and **high GPU parallelism**, shows strong results on both natural images [2, 3] and Whole Slide Images (WSI) [4].
- **Limitation #1**: **Spatial discrepancy** in **all** current mamba variants [2-4]. They are **inherently 1D** as they flatten 2D images into 1D sequences, losing spatial context.



Spatial discrepancy

Spatial continuity

$$h^{1D}_{3,3} = \boxed{\bar{A}^{8}}\bar{B}x_{1,1} + \bar{A}^{7}\bar{B}x_{1,2} + \bar{A}^{6}\bar{B}x_{1,3} + \bar{A}^{5}\bar{B}x_{2,1} + \bar{A}^{4}\bar{B}x_{2,2} + \bar{A}^{3}\bar{B}x_{2,3} + \bar{A}^{2}\bar{B}x_{3,1} + \bar{A}^{1}\bar{B}x_{3,2} + \bar{A}^{0}\bar{B}x_{3,3}$$

Large order leads to **forgetting**

$$h^{2D}_{3,3} = \boxed{\bar{A}^{4}}\bar{B}x_{1,1} + \bar{A}^{3}\bar{B}x_{1,2} + \bar{A}^{2}\bar{B}x_{1,3} + \bar{A}^{3}\bar{B}x_{2,1} + \bar{A}^{2}\bar{B}x_{2,2} + \bar{A}^{1}\bar{B}x_{2,3} + \bar{A}^{2}\bar{B}x_{3,1} + \bar{A}^{1}\bar{B}x_{3,2} + \bar{A}^{0}\bar{B}x_{3,3}$$

Manhattan distance (2D)

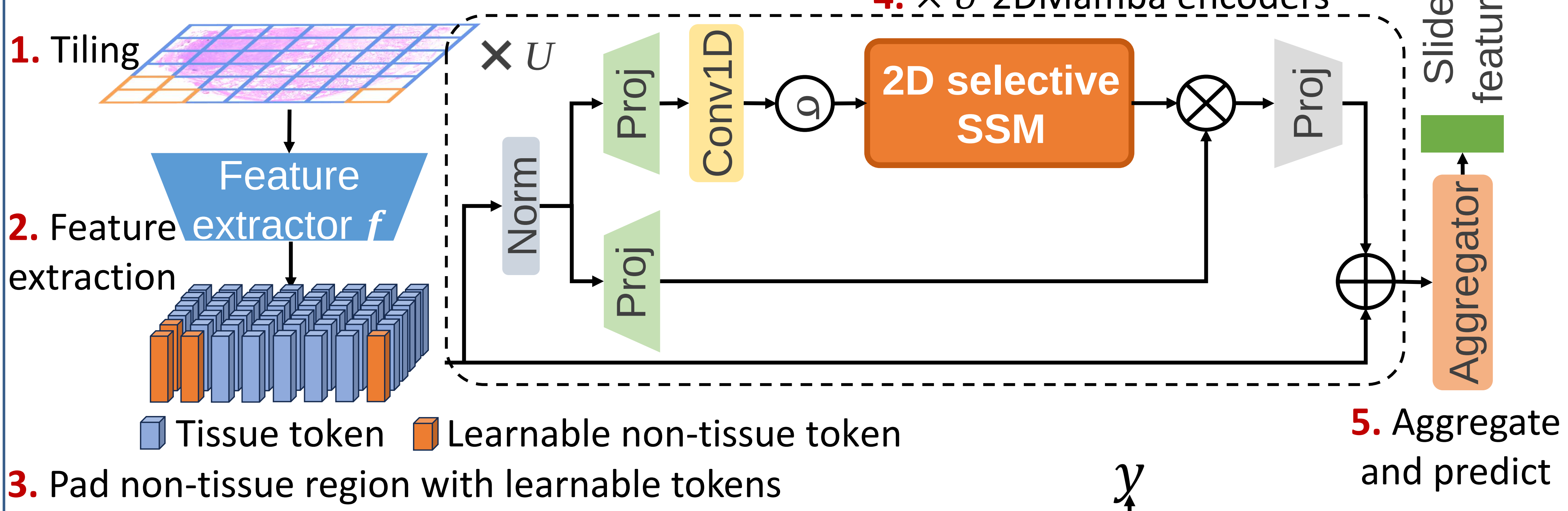
- **Limitation #2**: **Speed** in existing 2D SSM methods [5]. They still lack an efficient **parallel algorithm** because of their formulation and thus **slow**.

Our contributions:

- A novel **2D SSM** architecture that directly scans a 2D image without first flattening it into a 1D sequence, maintaining the **2D structure**.
- A fast **hardware-aware 2D** CUDA operator to extend the 1D Mamba **parallelism** into 2D.

Method

- **Overall framework:**



1. Tiling

2. Feature extraction

Feature extractor f

Tissue token

Learnable non-tissue token

3. Pad non-tissue region with learnable tokens

4. $\times\ U$ 2DMamba encoders

$\times\ U$

Norm

Proj

Conv1D

σ

2D selective SSM

Proj

Proj

Slide feature

Aggregator

5. Aggregate and predict

- **2D selective SSM:**

1. Horizontal scan:

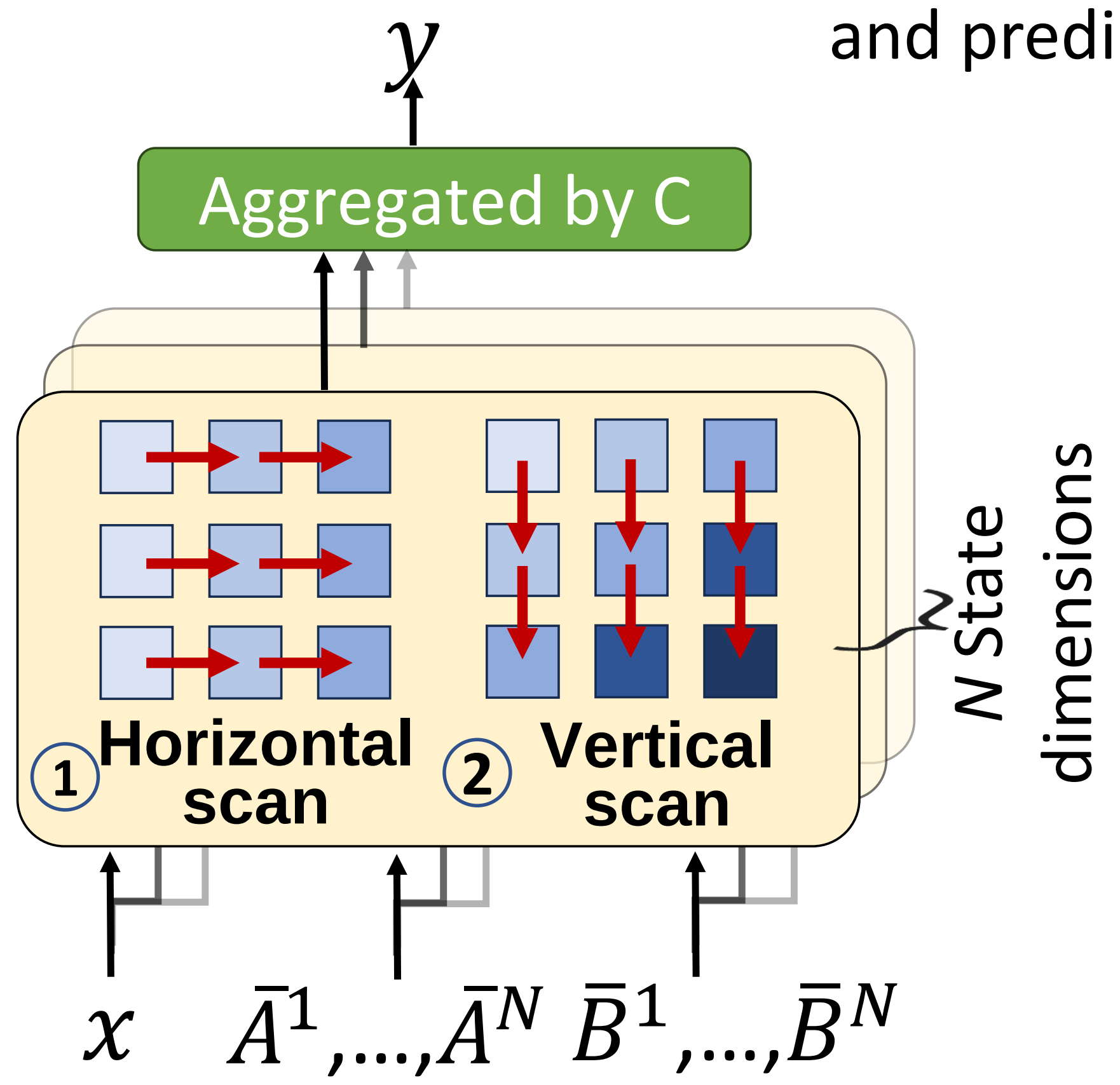
$$h^{hor}_{i,j} = \bar{A}_{i,j} h^{hor}_{i,j-1} + \bar{B}_{i,j} x_{i,j}$$

2. Vertical scan (without duplicated $\bar{B}_{i,j}$):

$$h_{i,j} = \bar{A}_{i,j} h_{i-1,j} + h^{hor}_{i,j}$$

3. Aggregated by C:

$$y_{i,j} = C h_{i,j}$$



y

Aggregated by C

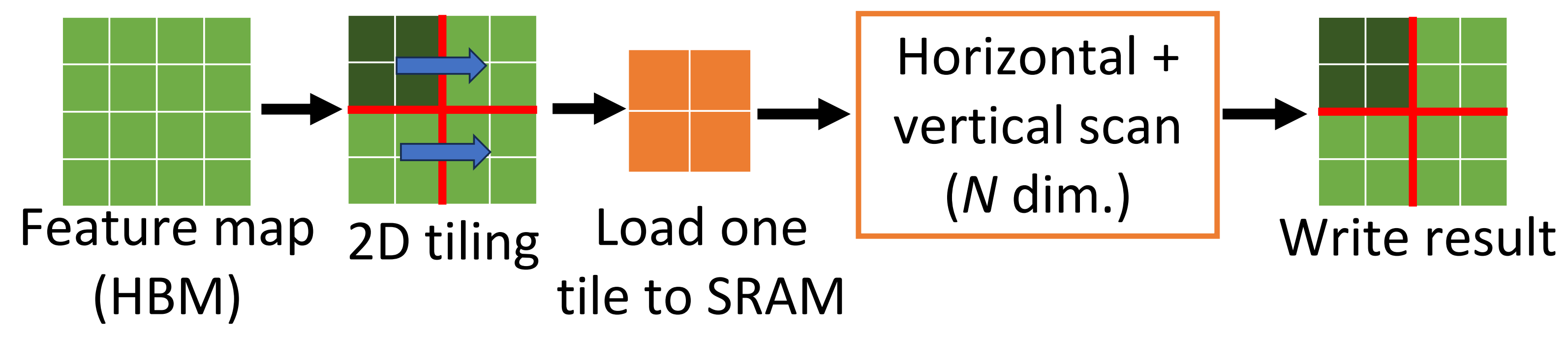
① Horizontal scan

② Vertical scan

N State dimensions

x $\bar{A}^1,\ldots,\bar{A}^N$ $\bar{B}^1,\ldots,\bar{B}^N$

- **Hardware-Aware 2D Selective Scan:**
  - Tile the feature map into **2D blocks**, scan each block in two directions, removing the naïve N× memory blow-up and delivers a **10 × speed-up**.



Feature map (HBM)

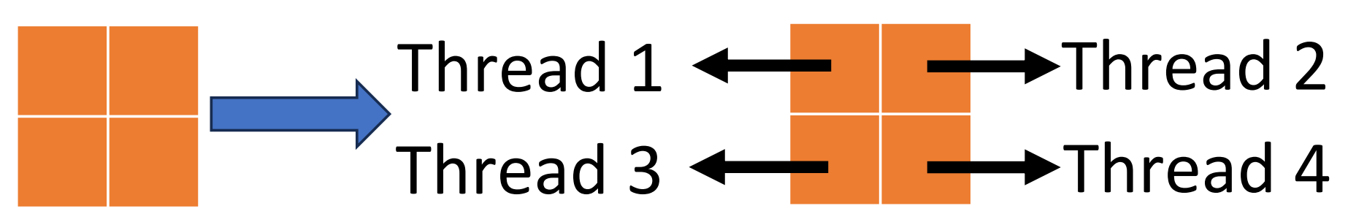
2D tiling

Load one tile to SRAM

Horizontal + vertical scan (N dim.)

Write result

  - Our SegmentedBlockScan enables scanning multiple rows and columns **in parallel**.

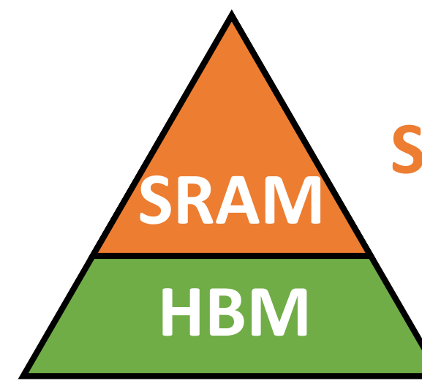


Thread 1

Thread 2

Thread 3

Thread 4



SRAM: 19TB/s

HBM: 1.5TB/s

Experiments

- **WSI benchmarks on 10 datasets**

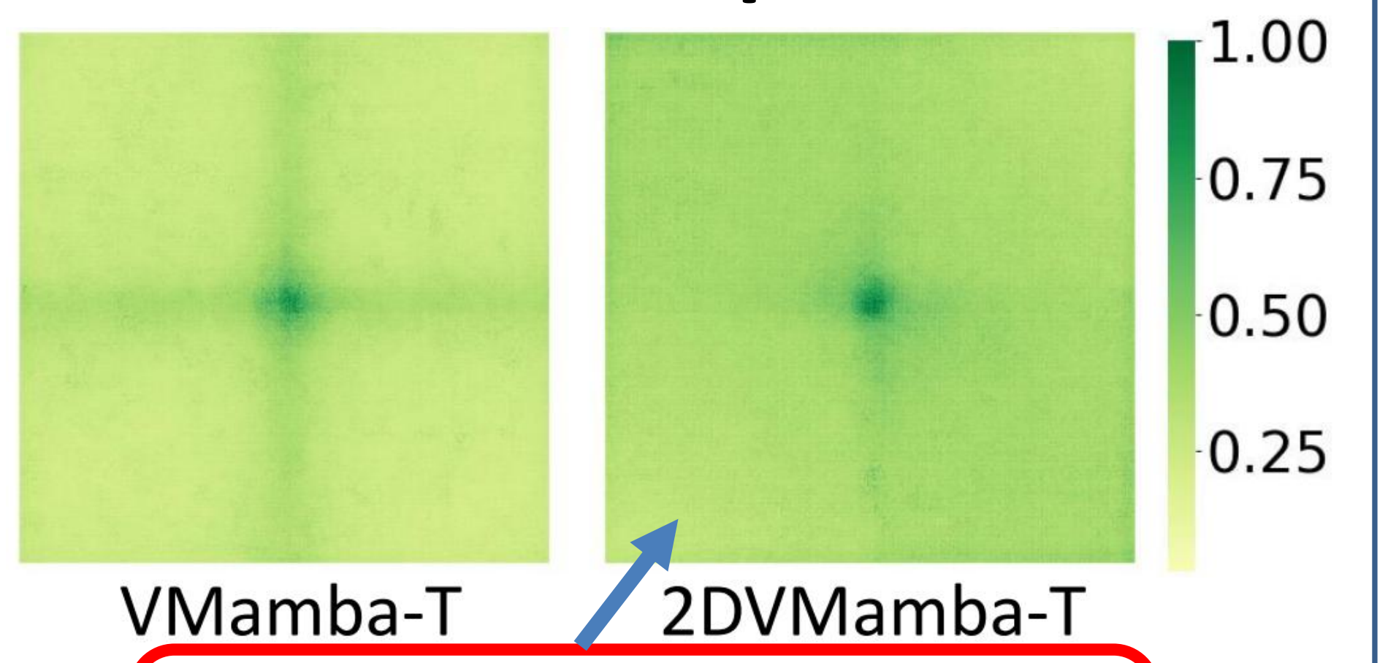
Method	WSI classification (Accuracy)					Survival analysis (C-index)				
	BRCAS	DHMC	PANDA	NSCLC	BRCA	KIRC	KIRP	LUAD	STAD	UCEC
DTFD-MIL	0.701	0.871	0.470	0.874	0.927	0.727	0.793	0.602	0.617	0.746
TransMIL	0.692	0.807	0.464	**0.885**	0.938	0.694	0.732	0.614	0.598	0.700
S4MIL	0.662	0.864	0.505	**0.885**	**0.946**	0.723	0.791	0.595	0.600	0.746
MmabaMIL	0.738	0.855	0.468	0.876	0.933	0.710	0.782	0.595	0.624	0.742
SRMambaMIL	0.738	0.859	0.471	**0.885**	0.931	0.718	0.742	0.588	0.613	0.740
2DMmabaMIL	**0.752**	**0.893**	**0.508**	**0.885**	**0.946**	**0.731**	**0.803**	**0.620**	**0.643**	**0.754**

2DMamba outperforms the SOTA methods

- **Natural image benchmarks**

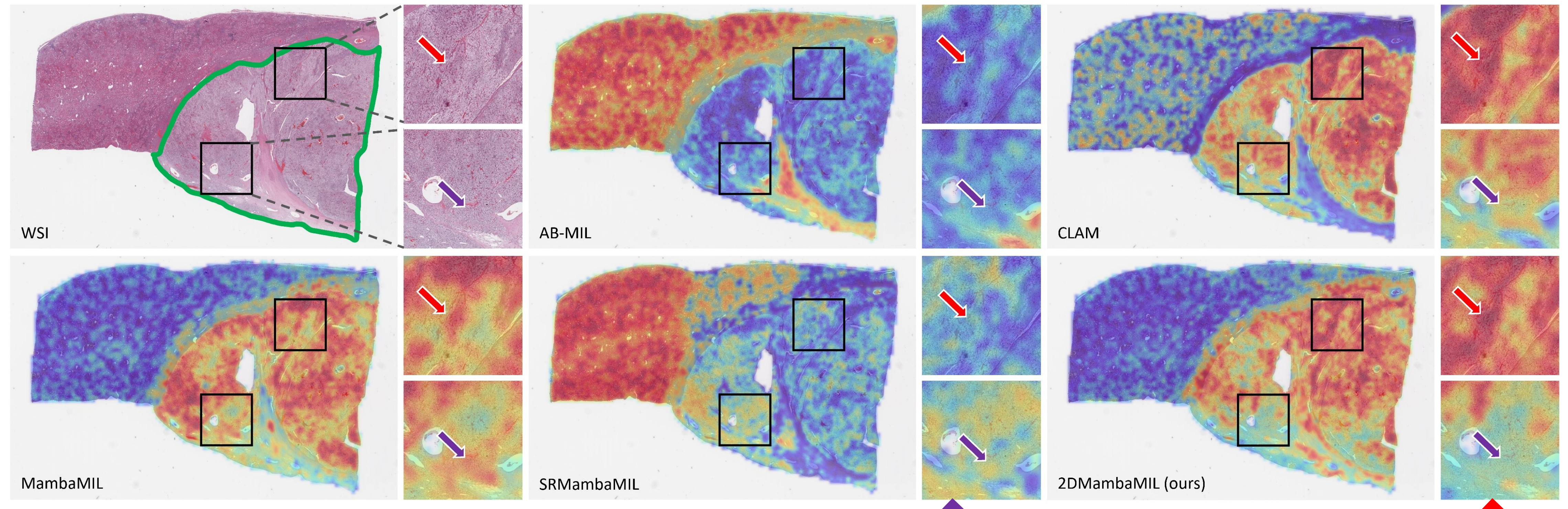
Method	ImagNet-1K	ADE20K	
	Top-1 Acc.	mIoU (SS)	mIoU (MS)
Vim-S	80.3%	44.9	-
EfficientVMamba-B	81.8%	46.5	47.3
LocalVMamba-T	82.7%	47.9	49.1
VMamba-T	82.6%	47.9	48.8
2DVMamba-T	**82.8%**	**48.6**	**49.3**

- **Effective Receptive Fields**



A more global pattern without cross-signal

- **Qualitative Evaluation on WSI**



Focusing on critical survival related regions ↘, not less related ones ↘

[1] Gu, Albert et.al. "Mamba: Linear-time sequence modeling with selective state spaces." arXiv preprint arXiv:2312.00752 (2023)
[2] Lianghui Zhu et.al. "Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model." ICML 2024
[3] Liu, Yue, et al. "Vmamba: Visual state space model." Advances in neural information processing systems 37 (2024): 103031-103063.
[4] Yang, Shu et.al. "Mambamil: Enhancing long sequence modeling with sequence reordering in computational pathology." MICCAI. Cham: Springer Nature Switzerland, 2024.
[5] Baron, Ethan et.al. "2-D SSM: A General Spatial Layer for Visual Transformers." arXiv preprint arXiv:2306.06635 (2023).

Acknowledgements: This work was partially supported by USA NSF grants IIS-2123920 [D.S], IIS-2212046 [D.S], IIS-1715985 [H.Q], IIS-1812606 [H.Q], the Canadian Cancer Society Breakthrough Grant [V.Q.H.T], FRQS-CRS-J1 [V.Q.H.T], NSERC-DG RGPIN-2022-05378 [M.S.H] and Amazon Research Award [M.S.H].