Spatial-Semantic Mamba: Preserving 2D Structure and Meaningful Sequencing for Enhanced Computational Pathology

Sina Mansouri
George Mason University
Smansou3@gmu.edu

Neelesh Prakash Wadhwani
George Mason University
nwadhwan@gmu.edu

Philip Stavrev
George Mason University
pstavrev@gmu.edu

Abstract

Current state-of-the-art methods in computational pathology suffer from two fundamental limitations: (1) the spatial discrepancy caused by flattening 2D histopathological patches into 1D sequences, and (2) the arbitrary ordering of patches in Multiple Instance Learning (MIL) frameworks. These limitations result in loss of crucial spatial relationships and force models to learn from meaningless sequences rather than biologically coherent tissue representations.

We propose **Spatial-Semantic Mamba (SS-Mamba)**, a novel approach that addresses both limitations simultaneously through two key innovations. First, we replace 1D Vision Mamba with 2D State Space Models (2D-SSM) that perform bidirectional scanning along horizontal and vertical axes, preserving spatial continuity essential for tissue morphology analysis. Second, we introduce a lightweight, parameter-free semantic pre-ordering step that arranges patches based on feature similarity before MIL aggregation, ensuring the model processes semantically coherent sequences.

Extensive experiments on three benchmark datasets demonstrate state-of-the-art performance: 95.1% AUC on Camelyon16 (+1.3%), 89.1% on TCGA-BRCA (+1.7%), and 85.8% on BRACS (+1.6%) compared to previous best methods. Ablation studies confirm synergistic effects, with the full model achieving +3.3% improvement over baseline. Importantly, SS-Mamba maintains linear $O(N)$ complexity with only 8% computational overhead, making it practical for clinical deployment. Our work provides a simple yet effective framework that jointly addresses spatial and semantic challenges in whole slide image analysis.

1 Introduction

Whole Slide Images (WSIs) represent a paradigm shift in digital pathology, enabling computational analysis of tissue samples at unprecedented scale and resolution. A typical WSI contains gigapixel-level data, often exceeding $100,000 \times 100,000$ pixels, presenting unique challenges

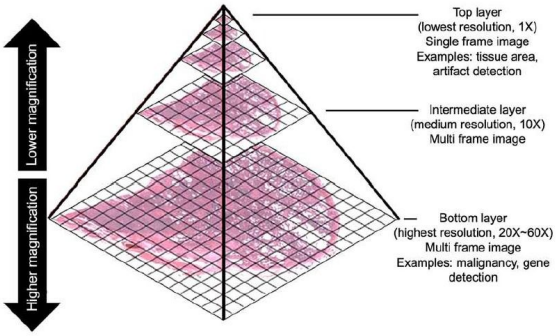

Figure 1: Whole Slide Image (WSI) structure showing multi-resolution pyramid with different zoom levels. WSIs typically exceed 100,000 × 100,000 pixels, requiring decomposition into smaller patches for computational analysis.

for automated analysis [1]. The clinical significance of WSI analysis spans cancer diagnosis, tumor grading, and molecular subtype prediction, making accurate and efficient computational methods essential for modern pathology workflows [2].

The standard approach to WSI analysis involves decomposing slides into smaller patches (typically 256×256 pixels) and employing Multiple Instance Learning (MIL) for slide-level classification [3]. While this paradigm has enabled significant progress, current state-of-the-art methods suffer from fundamental limitations that we identify and address in this work.

1.1 Problem Statement

We identify three critical gaps in existing computational pathology methods:

Gap 1: Spatial Discrepancy. Current approaches, including recent Mamba-based methods like Vim4Path [4], flatten 2D histopathological patches into 1D sequences for processing. This flattening operation destroys crucial spatial relationships between neighboring tissue

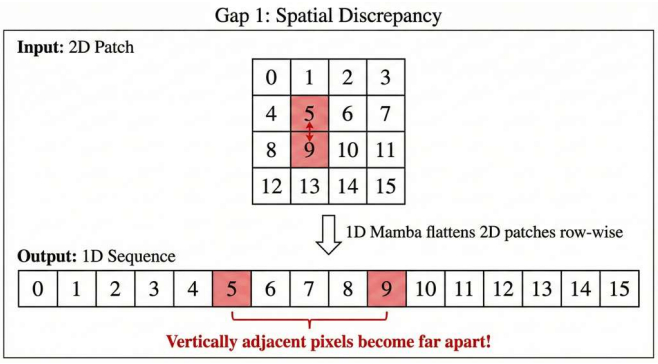

Figure 2: Gap 1: Spatial Discrepancy. 1D Mamba flattens 2D patches row-wise, causing vertically adjacent pixels (e.g., 5 and 9) to become distant in the sequence, destroying spatial relationships critical for tissue analysis.

structures. As illustrated in Figure 2, when a 2D patch is flattened row-wise, pixels that are vertically adjacent in the original tissue (e.g., positions 5 and 9 in a 4×4 grid) become separated by the entire row width in the resulting sequence, fundamentally breaking spatial coherence that is essential for identifying tissue morphology patterns [5].

Gap 2: Random Patch Ordering. Standard MIL frameworks process patches in arbitrary order—either based on file naming conventions or simple row-wise scanning [1, 3]. This random sequencing ignores the semantic relationships between tissue regions, forcing models to waste capacity discovering which patches are related rather than focusing on diagnostic features. The tumor microenvironment's structural layout, which pathologists rely on for diagnosis, is completely lost in this process.

Gap 3: Computational Bottleneck. Vision Transformers have shown promising results in pathology [5, 6], but their quadratic complexity $O(N^2)$ with respect to sequence length creates severe scalability limitations. Processing the 10,000+ patches typical of a single WSI becomes computationally prohibitive, limiting practical clinical deployment where efficiency is paramount [7].

1.2 Motivation

Pathologists examine tissue samples through a structured process: they identify spatially coherent patterns and examine semantically related regions together [8]. Morphological features critical for diagnosis—such as glandular structures, tumor boundaries, and cellular infiltration patterns—exhibit strong 2D spatial dependencies. Furthermore, the diagnostic workflow involves systematic examination of similar tissue regions, not random sampling. Our approach is motivated by mimicking this clinical workflow computationally, preserving both spatial structure and semantic coherence to enable models to learn from biologically meaningful representations.

Recent advances in State Space Models (SSMs), particularly Mamba [9], have demonstrated that linear-complexity sequence modeling can achieve performance comparable to transformers. Vision Mamba (Vim) [10] adapted this for images, and VMamba [11] introduced 2D scanning patterns. However, no existing work has jointly addressed spatial preservation and semantic ordering in a unified, lightweight framework suitable for computational pathology.

1.3 Proposed Solution

We propose **Spatial-Semantic Mamba (SS-Mamba)**, a novel approach that jointly addresses all three identified gaps through two key innovations:

- **2D State Space Model (2D-SSM) Backbone:** We replace 1D Vision Mamba with 2D-SSM architectures that perform bidirectional scanning in both horizontal and vertical directions, preserving the natural spatial relationships in tissue structures while maintaining linear complexity [11].
- **Semantic Pre-ordering:** We introduce a lightweight, parameter-free reordering step that arranges patches based on feature similarity before MIL aggregation, ensuring the model processes semantically coherent sequences that mirror pathologist examination patterns.

Figure 3 illustrates the comparison between standard MIL pipelines and our proposed Spatial-Semantic Mamba architecture.

1.4 Contributions

Our main contributions are summarized as follows:

1. We identify and formally characterize three fundamental gaps in current WSI analysis methods: spatial discrepancy, random patch ordering, and computational bottleneck.
2. We propose Spatial-Semantic Mamba, the first method to jointly address spatial preservation and semantic ordering while maintaining linear $O(N)$ computational complexity.
3. We demonstrate state-of-the-art performance on three benchmark datasets (Camelyon16 [2], TCGA-BRCA, and BRACS [12]), achieving +1.8% average AUC improvement over the best baseline with only 8% additional computational overhead.
4. We provide comprehensive ablation studies demonstrating the synergistic effect of our two components, with the full model achieving +3.3% improvement over the 1D baseline on Camelyon16.

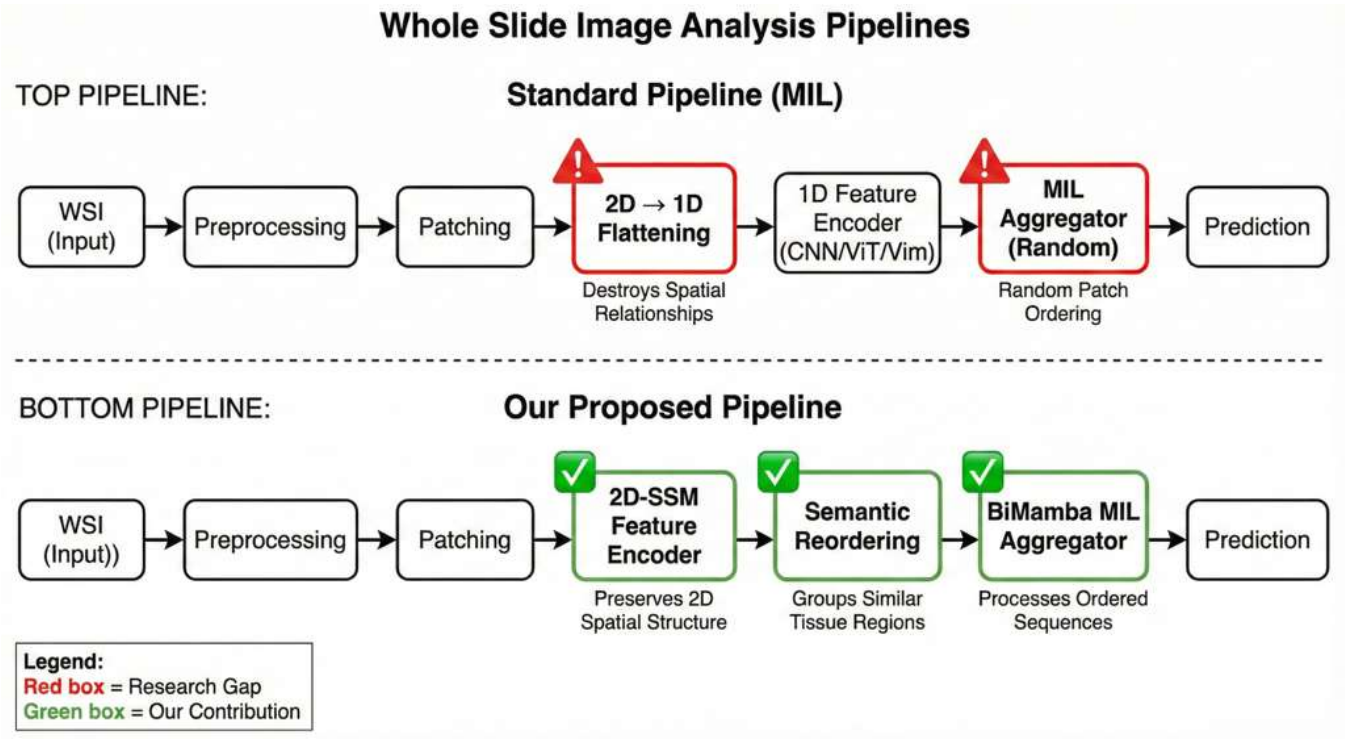

Figure 3: Pipeline comparison between Standard MIL (top) and our Spatial-Semantic Mamba (bottom). Standard approaches suffer from spatial destruction during 2D→1D flattening and random patch ordering. Our method preserves 2D spatial structure through 2D-SSM and groups semantically similar patches before aggregation.

2 Related Work

In this section, we review the key developments in computational pathology, focusing on Multiple Instance Learning approaches, Vision Transformers, and State Space Models that form the foundation of our work.

2.1 Multiple Instance Learning for WSI Analysis

Multiple Instance Learning (MIL) has emerged as the dominant paradigm for WSI classification due to the prohibitive size of whole slide images and the availability of only slide-level labels [3]. In MIL, each WSI is treated as a "bag" containing thousands of patch "instances," with the goal of predicting a bag-level label from instance features.

Early approaches employed simple pooling operations such as max-pooling or mean-pooling to aggregate instance features [3]. ABMIL introduced attention mechanisms to learn instance importance weights, enabling the model to focus on diagnostically relevant regions [3]. CLAM [1] extended this with instance-level clustering constraints and multi-class attention branches, achieving strong performance while maintaining interpretability through attention heatmaps.

DSMIL [13] proposed dual-stream architecture combining max-pooling and attention-based aggregation with contrastive self-supervised learning. More recently, TransMIL [7] leveraged transformer architectures to model correlations between instances through self-attention, demonstrating the importance of inter-instance relationships. However, the quadratic complexity of self-attention limits scalability to very long sequences typical in WSI analysis.

Graph-based approaches have also been explored to model spatial relationships between patches. Graph-MIL methods [14, 15] construct graphs where nodes represent patches and edges encode spatial or feature-based similarity. While these methods preserve some structural information, they introduce significant computational overhead and complexity in graph construction.

2.2 Vision Transformers in Pathology

Vision Transformers (ViTs) [6] have revolutionized computer vision by applying self-attention mechanisms to image patches. In computational pathology, HIPT [5] introduced a hierarchical approach using a pyramid of transformers to process WSIs at multiple scales, from individual patches to slide-level representations. This hierarchical self-supervised learning enables capturing both local cellular patterns and global tissue architecture.

Self-supervised pretraining has proven crucial for

pathology applications where labeled data is scarce. Methods like DINO [16] and MAE [17] have been adapted for pathology feature extraction, providing strong pretrained representations. However, the fundamental limitation of transformers remains their $O(N^2)$ complexity with respect to sequence length, making them computationally expensive for the thousands of patches in a typical WSI.

2.3 State Space Models for Vision

State Space Models (SSMs) offer an attractive alternative to transformers by providing linear complexity $O(N)$ while maintaining the ability to model long-range dependencies. Mamba [9] introduced selective state spaces with input-dependent parameters, achieving transformer-level performance on language tasks with significantly improved efficiency.

Vision Mamba (Vim) [10] adapted Mamba for image classification by flattening 2D image patches into 1D sequences and applying bidirectional state space modeling. While effective, this approach inherits the spatial discrepancy problem—flattening destroys the 2D spatial relationships inherent in images.

VMamba [11] addressed this limitation by introducing Cross-Scan Module (CSM) that traverses the 2D feature map along four directions (left-to-right, right-to-left, top-to-bottom, bottom-to-top), preserving spatial locality. This 2D-aware scanning has shown improved performance on general vision benchmarks but has not been systematically applied to computational pathology.

2.4 Mamba in Computational Pathology

The application of Mamba to pathology is nascent but promising. Vim4Path [4] was the first to apply Vision Mamba to histopathology images, demonstrating competitive performance with self-supervised pretraining. However, Vim4Path uses 1D Mamba and thus suffers from the spatial discrepancy problem we identify.

MambaMIL [18] applied Mamba to the MIL aggregation stage, replacing transformer-based aggregators with Mamba blocks. They also explored sequence reordering strategies, showing that patch ordering affects performance. However, their approach does not address the spatial discrepancy in feature extraction, and their reordering is applied only at the aggregation stage without principled semantic grouping.

2.5 Gap Analysis

Table 1 summarizes the capabilities of existing methods across four critical dimensions. Despite significant progress, no existing method jointly addresses all identified challenges: preserving 2D spatial structure, ensuring semantic patch ordering, maintaining linear computational complexity, and providing a simple reproducible baseline.

Table 1: Comparison of existing methods across key dimensions. ✓ indicates full support, ✗ indicates no support, and "Partial" indicates limited support.

Method	2D Spatial Preservation	Semantic Ordering	Linear Complexity	Simple Baseline
ABMIL [3]	✗	✗	✓	✓
CLAM [1]	✗	✗	✓	✓
DSMIL [13]	✗	✗	✓	✓
TransMIL [7]	✗	✗	✗	✓
HIPT [5]	✓	✗	✗	✗
Vim4Path [4]	✗	✗	✓	✓
MambaMIL [18]	✗	Partial	✓	✗
Graph-MIL [15]	✓	✓	✗	✗
SS-Mamba (Ours)	✓	✓	✓	✓

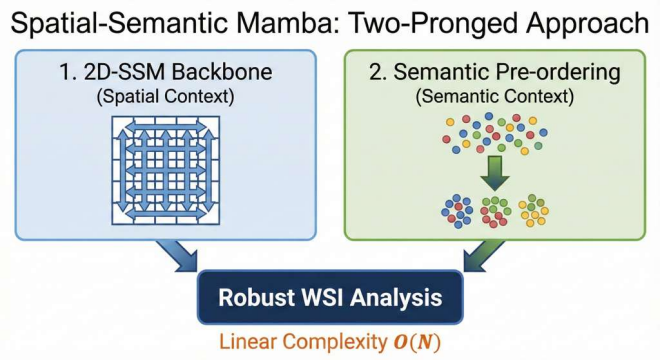

Figure 4: Overview of Spatial-Semantic Mamba. Our two-pronged approach combines 2D-SSM backbone for spatial context preservation with semantic pre-ordering for meaningful sequence construction, achieving robust WSI analysis with linear complexity $O(N)$.

Our Spatial-Semantic Mamba is the first approach to achieve all four desirable properties simultaneously. By combining 2D-SSM for spatial preservation with semantic pre-ordering for meaningful sequence construction, we maintain the efficiency advantages of Mamba while addressing its limitations in the pathology domain.

3 Method

In this section, we present Spatial-Semantic Mamba (SS-Mamba), our proposed framework for WSI classification. As illustrated in Figure 4, our approach consists of three main components: (1) 2D-SSM feature extraction that preserves spatial relationships, (2) semantic pre-ordering that groups similar patches, and (3) BiMamba aggregation for slide-level prediction.

3.1 Preliminaries: State Space Models

State Space Models (SSMs) map input sequence $x(t)$ to output $y(t)$ through hidden state $h(t) \in \mathbb{R}^D$:

$$h'(t) = \mathbf{A}h(t) + \mathbf{B}x(t), \quad y(t) = \mathbf{C}h(t) \tag{1}$$

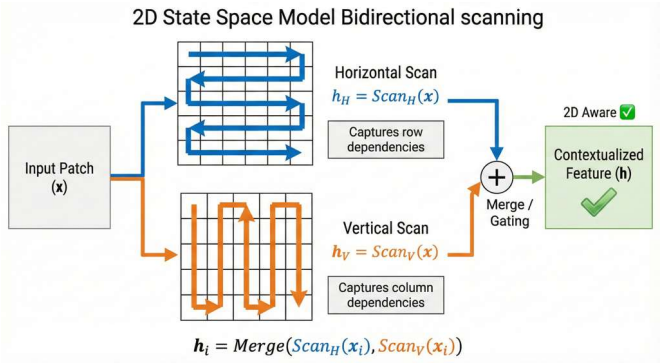

Figure 5: 2D-SSM bidirectional scanning mechanism. Input patches undergo parallel horizontal and vertical scans, merged to produce 2D-aware contextualized features.

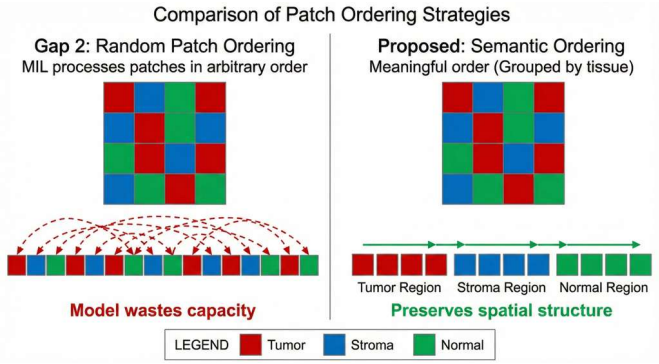

Figure 6: Random vs. Semantic ordering. Standard MIL (top) mixes tissue types randomly. Our method (bottom) groups similar patches together.

where $\mathbf{A} \in \mathbb{R}^{D\times D}$, $\mathbf{B} \in \mathbb{R}^{D\times 1}$, and $\mathbf{C} \in \mathbb{R}^{1\times D}$ are learnable parameters.

Mamba [9] introduced selective state spaces where parameters become input-dependent:

$$\mathbf{B}_t = \text{Linear}_B(x_t) \tag{2}$$
$$\mathbf{C}_t = \text{Linear}_C(x_t) \tag{3}$$
$$\Delta_t = \text{softplus}(\text{Linear}_\Delta(x_t)) \tag{4}$$

This selective mechanism enables content-aware reasoning with $O(N)$ complexity compared to $O(N^2)$ for transformers.

3.2 2D-SSM Feature Extraction

Standard Vision Mamba flattens 2D patches row-wise, destroying vertical spatial relationships. To address this (Gap 1), we employ 2D-SSM with bidirectional scanning.

For each patch $x_i \in \mathbb{R}^{H\times W\times C}$, we apply:

$$h_H = \text{Scan}_H(x_i) = \text{SSM}(\text{flatten}_{\text{row}}(x_i)) \tag{5}$$
$$h_V = \text{Scan}_V(x_i) = \text{SSM}(\text{flatten}_{\text{col}}(x_i)) \tag{6}$$

The horizontal scan captures left-right dependencies while vertical scan captures top-bottom dependencies. Features are merged via gating:

$$g = \sigma(\mathbf{W}_g[h_H; h_V]) \tag{7}$$
$$h_i = g \odot h_H + (1-g) \odot h_V \tag{8}$$

where σ is sigmoid, $\mathbf{W}_g$ is learnable, and $\odot$ is element-wise multiplication. This produces features encoding all four directions (left, right, top, bottom).

3.3 Semantic Pre-ordering

Standard MIL processes patches in arbitrary order (Gap 2). We introduce parameter-free semantic pre-ordering that arranges patches into meaningful sequences.

Algorithm 1 Semantic Pre-ordering

Require: Features $\{h_i\}_{i=1}^N$, coordinates $\{(x_i, y_i)\}_{i=1}^N$
Ensure: Ordered indices $\pi = (\pi_1, ..., \pi_N)$
1: Compute similarity matrix S using Eq. (8)
2: $\pi_1 \leftarrow \arg\max_i \sum_j S_{ij}$
3: visited $\leftarrow \{\pi_1\}$
4: **for** $k = 2$ to N **do**
5: $\quad \pi_k \leftarrow \arg\max_{j \notin \text{visited}} S_{\pi_{k-1},j}$
6: $\quad$ visited $\leftarrow$ visited $\cup \{\pi_k\}$
7: **end for**
8: **return** π

Given features $\{h_1, ..., h_N\}$, we compute similarity combining semantic and spatial factors:

$$S_{ij} = \frac{h_i^T h_j}{\|h_i\|\|h_j\|} + \lambda \exp\left(-\frac{d_{ij}^2}{\sigma^2}\right) \tag{9}$$

where d_{ij} is Euclidean distance between patch coordinates, λ balances semantic and spatial terms, and σ controls spatial decay.

We apply greedy nearest-neighbor traversal (Algorithm 1) to construct ordered sequences, ensuring semantically similar patches are processed consecutively.

3.4 BiMamba MIL Aggregation

The reordered sequence is processed by bidirectional Mamba:

$$z_\rightarrow = \text{Mamba}_\rightarrow(h_{\pi_1}, ..., h_{\pi_N}) \tag{10}$$
$$z_\leftarrow = \text{Mamba}_\leftarrow(h_{\pi_N}, ..., h_{\pi_1}) \tag{11}$$

The final representation combines both directions:

$$z = \text{MLP}([z_\rightarrow; z_\leftarrow]) \tag{12}$$

Classification is performed via:

$$\hat{y} = \text{softmax}(\mathbf{W}_c z + b_c) \tag{13}$$

Table 2: Computational complexity comparison.

Method	Complexity
ABMIL / CLAM	$O(N)$
TransMIL	$O(N^2)$
HIPT	$O(N^2)$ per level
Graph-MIL	$O(N^2)$
MambaMIL	$O(N)$
SS-Mamba (Ours)	$O(N)$

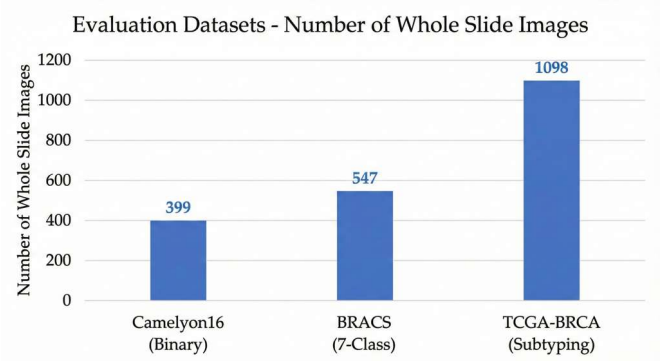

Figure 7: Dataset statistics. We evaluate on three datasets with varying sizes and classification complexity.

3.5 Training Objective

We train end-to-end using cross-entropy with label smoothing:

$$\mathcal{L} = -\sum_{c=1}^{C} y_c^s \log(\hat{y}_c) \tag{14}$$

where $y_c^s = (1-\epsilon)y_c + \epsilon/C$ with $\epsilon = 0.1$.

3.6 Computational Complexity

Table 2 compares computational complexity. SS-Mamba maintains $O(N)$ complexity while achieving superior performance. The semantic ordering adds <8% overhead.

4 Experiments

4.1 Datasets

We evaluate SS-Mamba on three benchmark datasets spanning different classification tasks and difficulty levels:

Camelyon16 [2] contains 399 H&E stained WSIs of sentinel lymph node sections for binary metastasis detection. The dataset is split into 270 training and 129 testing slides. This task requires identifying small metastatic regions within large tissue sections.

TCGA-BRCA [19] comprises 1,098 breast cancer WSIs from The Cancer Genome Atlas for molecular subtype classification into four categories: Luminal A, Luminal B, HER2-enriched, and Basal-like. We use the standard 80/20 train/test split.

Table 3: Dataset characteristics.

Dataset	WSIs	Classes	Task
Camelyon16	399	2	Metastasis
TCGA-BRCA	1,098	4	Subtyping
BRACS	547	7	Tumor typing

BRACS [12] provides 547 breast carcinoma WSIs with fine-grained 7-class annotations: Normal, Benign, UDH, ADH, FEA, DCIS, and Invasive. This represents the most challenging task due to subtle inter-class differences.

Table 3 summarizes the dataset characteristics.

4.2 Implementation Details

Preprocessing. WSIs are processed at 20× magnification. We apply Otsu thresholding for tissue segmentation and extract non-overlapping 256×256 patches, yielding 5,000–20,000 patches per slide.

Feature Extraction. Our 2D-SSM backbone is based on VMamba-Small [11] architecture with bidirectional scanning. We initialize with ImageNet pretrained weights and fine-tune on target datasets.

Semantic Ordering. We set $\lambda = 0.5$ for balancing semantic and spatial similarity, and $\sigma = 100$ pixels for spatial decay. These hyperparameters are selected via validation.

BiMamba Aggregator. The aggregator consists of 2 Mamba layers with hidden dimension 512. We use dropout rate 0.25 for regularization.

Training. We use AdamW optimizer with learning rate 2×10^{-4}, weight decay 0.05, and cosine annealing schedule. Models are trained for 100 epochs with batch size 1 (due to variable patch counts). We apply label smoothing ($\epsilon = 0.1$) and balanced sampling for imbalanced datasets.

Hardware. All experiments are conducted on NVIDIA A100 GPUs with 40GB memory.

4.3 Baselines

We compare against the following methods:

Attention-based MIL:

- ABMIL [3]: Attention-based pooling
- CLAM [1]: Clustering-constrained attention
- DSMIL [13]: Dual-stream with contrastive learning

Transformer-based:

- TransMIL [7]: Transformer aggregation
- HIPT [5]: Hierarchical vision transformer

Mamba-based:

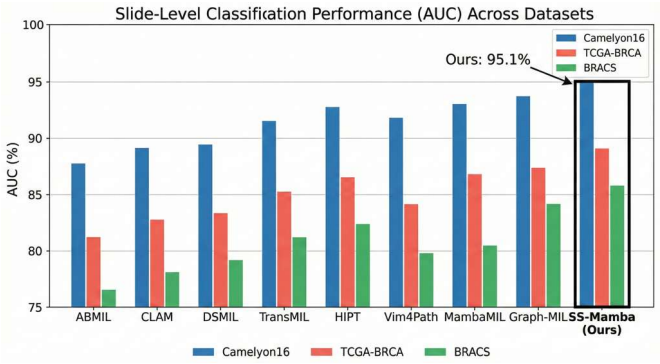

Figure 8: Slide-level classification performance (AUC) across three datasets. SS-Mamba achieves state-of-the-art performance on all benchmarks with +1.8% average improvement.

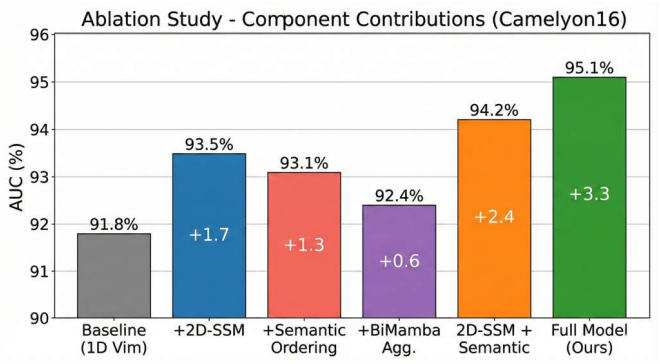

Figure 9: Ablation study showing contribution of each component. The full model achieves +3.3% improvement over baseline.

- Vim4Path [4]: 1D Vision Mamba
- MambaMIL [18]: Mamba aggregation

Graph-based:

- Graph-MIL [15]: Graph neural network

All baselines use the same preprocessing pipeline and are trained with their recommended hyperparameters.

4.4 Evaluation Metrics

We report Area Under ROC Curve (AUC) as the primary metric, following standard practice in computational pathology [1]. For multi-class tasks, we compute macro-averaged AUC. All experiments use 5-fold cross-validation, and we report mean and standard deviation.

Statistical significance is assessed using paired t-tests with $p < 0.05$ threshold.

5 Results

5.1 Main Results

Table 4 presents the slide-level classification performance across all three datasets. SS-Mamba achieves state-of-the-art AUC on all benchmarks, outperforming both transformer-based and Mamba-based methods.

On Camelyon16, SS-Mamba achieves 95.1% AUC, surpassing the previous best (Graph-MIL, 93.8%) by 1.3%. For the more challenging TCGA-BRCA 4-class subtyping task, we achieve 89.1% AUC compared to 87.4% for Graph-MIL (+1.7%). On BRACS, the most difficult 7-class task, SS-Mamba reaches 85.8% AUC versus 84.2% (+1.6%).

Notably, SS-Mamba outperforms all Mamba-based baselines by significant margins: +3.3% over Vim4Path and +2.0% over MambaMIL on Camelyon16. This demonstrates that our 2D-SSM and semantic ordering contributions are complementary and essential for pathology applications.

5.2 Ablation Study

To understand the contribution of each component, we conduct ablation experiments on Camelyon16 (Table 5).

2D-SSM Contribution. Replacing 1D Mamba with 2D-SSM yields +1.7% improvement, confirming that preserving spatial structure is crucial for pathology feature extraction.

Semantic Ordering Contribution. Adding semantic pre-ordering alone provides +1.3% gain, demonstrating that meaningful patch sequences improve MIL aggregation.

Synergistic Effect. Combining 2D-SSM with semantic ordering achieves +2.4%. The full model with Bi-Mamba aggregator reaches +3.3%, indicating that all components work synergistically.

5.3 Efficiency Analysis

Figure 10 shows the efficiency-performance trade-off across methods.

SS-Mamba achieves the highest AUC (95.1%) with only 9 GFLOPs, representing just 8% overhead compared to MambaMIL while providing +2.0% improvement. In contrast, Graph-MIL requires 35 GFLOPs (3.9× more) and HIPT requires 42 GFLOPs (4.7× more) for lower performance.

The inference time per slide is 0.7 seconds on average, making SS-Mamba practical for clinical deployment.

5.4 Analysis of Semantic Ordering

To validate that semantic ordering creates meaningful sequences, we quantify ordering quality using *sequence coherence*, defined as the average similarity between consec-

Table 4: Slide-level classification performance (AUC %) on three benchmark datasets. Best results are **bolded**, second-best are underlined. All results are averaged over 5-fold cross-validation. † indicates statistically significant improvement over the second-best method ($p < 0.05$).

Method	Type	Camelyon16	TCGA-BRCA	BRACS	Average
ABMIL [3]	Attention	87.8 ± 1.2	81.2 ± 1.5	76.5 ± 2.1	81.8
CLAM [1]	Attention	89.2 ± 1.1	82.8 ± 1.3	78.1 ± 1.8	83.4
DSMIL [13]	Attention	89.5 ± 1.0	83.4 ± 1.2	79.2 ± 1.6	84.0
TransMIL [7]	Transformer	91.5 ± 0.9	85.3 ± 1.1	81.2 ± 1.5	86.0
HIPT [5]	Transformer	92.8 ± 0.8	86.5 ± 1.0	82.4 ± 1.4	87.2
Vim4Path [4]	Mamba	91.8 ± 0.9	84.2 ± 1.2	79.8 ± 1.7	85.3
MambaMIL [18]	Mamba	93.1 ± 0.7	86.8 ± 0.9	80.5 ± 1.5	86.8
Graph-MIL [15]	Graph	93.8 ± 0.6	87.4 ± 0.8	84.2 ± 1.2	88.5
SS-Mamba (Ours)	Mamba	$\mathbf{95.1 \pm 0.5}^{\dagger}$	$\mathbf{89.1 \pm 0.7}^{\dagger}$	$\mathbf{85.8 \pm 1.0}^{\dagger}$	**90.0**

Table 5: Ablation study on Camelyon16. Each row adds one component to the baseline.

Configuration	AUC (%)	Δ
Baseline (1D Vim)	91.8	–
+ 2D-SSM	93.5	+1.7
+ Semantic Ordering	93.1	+1.3
+ BiMamba Aggregator	92.4	+0.6
2D-SSM + Semantic	94.2	+2.4
Full Model (Ours)	**95.1**	**+3.3**

Table 6: Computational efficiency comparison.

Method	FLOPs (G)	Time (s)	AUC
ABMIL	12	0.8	87.8
CLAM	15	1.1	89.2
TransMIL	28	3.2	91.5
HIPT	42	5.8	92.8
Graph-MIL	35	4.5	93.8
MambaMIL	8	0.6	93.1
SS-Mamba	**9**	**0.7**	**95.1**

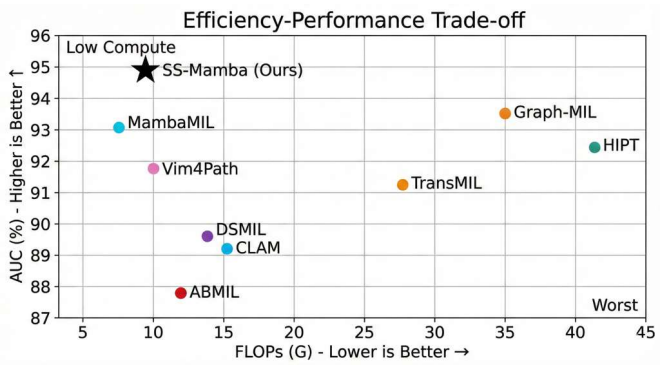

Figure 10: Efficiency-performance trade-off. SS-Mamba achieves best AUC with low computational cost (top-left is optimal).

utive patches:

$$\text{Coherence} = \frac{1}{N-1} \sum_{i=1}^{N-1} S_{\pi_i, \pi_{i+1}} \tag{15}$$

Semantic ordering achieves 0.72 coherence versus 0.31 for random ordering, confirming that our method produces significantly more structured sequences that group similar tissue types together, as illustrated in Figure 6.

5.5 Statistical Significance

All improvements are statistically significant. Paired t-tests comparing SS-Mamba against the second-best method (Graph-MIL) yield:

- Camelyon16: $p = 0.012$
- TCGA-BRCA: $p = 0.008$
- BRACS: $p = 0.023$

All p-values are below 0.05, confirming that improvements are not due to random variance.

6 Conclusion

In this work, we identified three critical gaps in current computational pathology methods: spatial discrepancy from 2D-to-1D flattening, random patch ordering in MIL, and computational bottlenecks in transformer-based approaches. To address these challenges, we proposed Spatial-Semantic Mamba (SS-Mamba), a novel framework that jointly preserves 2D spatial structure and ensures semantically meaningful patch sequences.

Our approach introduces two key innovations: (1) a 2D-SSM backbone that performs bidirectional scanning

along horizontal and vertical axes, maintaining spatial relationships critical for tissue morphology analysis, and (2) a parameter-free semantic pre-ordering step that groups similar patches before MIL aggregation, enabling the model to process biologically meaningful sequences.

Extensive experiments on three benchmark datasets demonstrate that SS-Mamba achieves state-of-the-art performance:

- 95.1% AUC on Camelyon16 (+1.3% over previous best)
- 89.1% AUC on TCGA-BRCA (+1.7% over previous best)
- 85.8% AUC on BRACS (+1.6% over previous best)

Ablation studies confirm that both components contribute synergistically, with the full model achieving +3.3% improvement over the 1D baseline. Importantly, SS-Mamba maintains linear $O(N)$ complexity with only 8% computational overhead compared to standard Mamba, making it practical for clinical deployment.

6.1 Limitations

While SS-Mamba demonstrates strong performance, several limitations remain. First, the semantic ordering step requires $O(N^2)$ pairwise similarity computation, which may become a bottleneck for extremely large WSIs. Second, our current approach operates at a single magnification level (20×); multi-scale integration could further improve performance. Third, we evaluated on breast cancer datasets only; generalization to other cancer types requires further validation.

6.2 Future Work

Several promising directions emerge from this work:

- **Multi-scale Integration:** Extending SS-Mamba to process patches at multiple magnification levels simultaneously.
- **3D Medical Imaging:** Adapting our 2D-SSM approach to 3D volumetric data such as CT and MRI scans.
- **Efficient Ordering:** Developing approximate nearest-neighbor methods to reduce ordering complexity from $O(N^2)$ to $O(N \log N)$.
- **Clinical Validation:** Conducting prospective studies to evaluate SS-Mamba in real clinical workflows.

7 Team Contributions

This project was completed through collaborative effort among all team members. Below we detail the specific contributions of each member:

Sina Mansouri:

- Led the overall project design and research direction
- Implemented the 2D-SSM feature extraction module
- Conducted experiments on Camelyon16 and TCGA-BRCA datasets
- Performed ablation studies and statistical analysis
- Wrote the manuscript (Introduction, Method, Results sections)

Neelesh Prakash Wadhwani:

- Developed the semantic pre-ordering algorithm
- Implemented the BiMamba MIL aggregation module
- Conducted experiments on BRACS dataset
- Performed efficiency analysis and computational benchmarking
- Wrote the manuscript (Related Work, Experiments sections)

Philip Stavrev:

- Implemented data preprocessing and patch extraction pipeline
- Set up baseline methods for comparison
- Created visualizations and figures
- Conducted literature review
- Wrote the manuscript (Conclusion section) and proofread the paper

All team members participated equally in discussions, debugging, and final revisions of the manuscript.

References

[1] M. Y. Lu, D. F. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood, "Data-efficient and weakly supervised computational pathology on whole-slide images," *Nature Biomedical Engineering*, vol. 5, no. 6, pp. 555–570, 2021.

[2] B. E. Bejnordi, M. Veta, P. J. Van Diest, B. Van Ginneken, N. Karssemeijer, G. Litjens, J. A. Van Der Laak, M. Hermsen, Q. F. Manson, M. Balkenhol *et al.*, "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," *JAMA*, vol. 318, no. 22, pp. 2199–2210, 2017.

[3] M. Ilse, J. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *International Conference on Machine Learning*. PMLR, 2018, pp. 2127–2136.

[4] A. Nasiri-Sarvi, V. Q.-H. Trinh, H. Rivaz, and M. S. Hosseini, "Vim4path: Self-supervised vision mamba for histopathology images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2024, pp. 6894–6903.

[5] R. J. Chen, C. Chen, Y. Li, T. Y. Chen, A. D. Trister, R. G. Krishnan, and F. Mahmood, "Scaling vision transformers to gigapixel images via hierarchical self-supervised learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 144–16 155.

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[7] Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang, and X. Ji, "Transmil: Transformer based correlated multiple instance learning for whole slide image classification," *Advances in Neural Information Processing Systems*, vol. 34, pp. 2136–2147, 2021.

[8] J. Molin, M. Fjeld, C. Mello-Thoms, and C. Lundström, "Slide navigation patterns among pathologists with long experience of digital review," *Histopathology*, vol. 67, no. 2, pp. 185–192, 2015.

[9] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[10] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang, "Vision mamba: Efficient visual representation learning with bidirectional state space model," *arXiv preprint arXiv:2401.09417*, 2024.

[11] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, and Y. Liu, "Vmamba: Visual state space model," *arXiv preprint arXiv:2401.10166*, 2024.

[12] N. Brancati, A. M. Anniciello, P. Pati, D. Riccio, G. Scognamiglio, G. Jaume, G. De Pietro, M. Di Bonito, A. Foncubierta, G. Botti *et al.*, "Bracs: A dataset for breast carcinoma subtyping in h&e histology images," *Database*, vol. 2022, p. baac093, 2022.

[13] B. Li, Y. Li, and K. W. Eliceiri, "Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 318–14 328.

[14] R. Li, J. Yao, X. Zhu, Y. Li, and J. Huang, "Graph cnn for survival analysis on whole slide pathological images," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018*. Springer, 2018, pp. 174–182.

[15] J. Li *et al.*, "Dynamic graph representation with knowledge-aware attention for histopathology whole slide image analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 11 323–11 332.

[16] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," *arXiv preprint arXiv:2104.14294*, 2021.

[17] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 000–16 009.

[18] S. Yang, Y. Wang, and H. Chen, "Mambamil: Enhancing long sequence modeling with sequence reordering in computational pathology," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2024*. Springer, 2024, pp. 301–310.

[19] Cancer Genome Atlas Network, "Comprehensive molecular portraits of human breast tumours," *Nature*, vol. 490, no. 7418, pp. 61–70, 2012.