Medical Image Analysis (2025)

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

SSMamba: A Self-Supervised Hybrid State Space Model for ROI Pathological Image Classification

Enhui Chai[a,*], Sicheng Chen[b,*], Tianyi Zhang[c], Xingyu Li[a], Tianxiang Cui[d,**]

[a] School of Computer Science, Northwest University, Xi'an 710127, China
[b] PuzzleLogic Pte Ltd, Singapore 229594, Singapore
[c] Department of Electrical & Computer Engineering, National University of Singapore, Singapore 117417, Singapore
[d] School of Computer Science, University of Nottingham Ningbo China, Ningbo 315100, China

ARTICLE INFO

Article history:

Keywords: Pathological ROI classification, Self-supervised Learning, State Space Model

ABSTRACT

Pathological diagnosis is essential for patient care, with region-of-interest (ROI) analysis serving as a critical pathological approach to extract localized cellular details that guide precise clinical decisions. Extracting robust visual representations from pathology ROI datasets is crucial yet remains challenging due to limited annotations and domain-specific complexities inherent to these focused image regions. While self-supervised learning (SSL) has shown promise in leveraging unlabeled data, existing approaches often fall short in addressing the unique characteristics of pathology ROI data from three aspects: (1) translation invariance, (2) local-global feature integration, and (3) domain shift. To address them, we present SSMamba, a novel hybrid SSL framework tailored for pathological ROI image classification. The key innovation lies in the synergistic integration of three corresponding domain-aware components: (1) To mitigate translation invariance, we designed Local Perception Residual (LPR) Module, which encodes spatially-aware local features by preserving fine-grained structural arrangements within ROIs; (2) To capture both localized patterns and global tissue structures across ROI regions, we proposed Directional Multi-scale (DMS) Module, which enables more holistic feature extraction across varying resolutions through directional multi-scale aggregation; (3) To improve generalization across scanners and institutions for ROI data, we proposed Mamba masked image modeling (MAMIM), a novel masked image modeling SSL pretraining method. By leveraging a unique decoder design to reconstruct partially masked inputs from pathological ROI datasets, this approach effectively promotes scanner-invariant pretraining. Across 7 public ROI datasets, SSMamba surpasses 6 state-of-the-art pathology foundation models, underscoring the necessity of architectures that explicitly capture pathological characteristics and showing that task-specific training—even with limited data—yields superior visual representations to generic ViT-style foundations.

1. Introduction

Pathology diagnosis is indispensable in clinical practice, as it relies on the detailed analysis of pathological images to enable accurate disease detection and inform precise treatment

*Equal contribution.
**Corresponding author: Tianxiang Cui (tianxiang.cui@nottingham.edu.cn)

planning (Srinidhi et al., 2021). Among the key tools for such analysis, Region-of-Interest (ROI) refers to specific subregions within pathological images that contain critical diagnostic information (e.g., abnormal cells or tissue structures), serving as the primary focus for detailed examination due to their relevance to disease characteristics. Traditional ROI analysis approaches largely depended on handcrafted features, which are often subjective and suffer from limited expressiveness (Madabhushi and Lee, 2016). In contrast, deep learning has shown superior representation learning capabilities (LeCun et al., 2015), but its success hinges on large-scale annotated datasets—resources that are particularly scarce and expensive in the pathology domain. To alleviate this dependency, transfer learning from natural image datasets such as ImageNet (Deng et al., 2009) has been widely adopted (Senousy et al., 2021). However, such pretraining strategies often fail to account for the domain-specific distribution and semantic gap between natural and pathological images, leading to suboptimal representations. To further improve the performance of the pretraining, self-supervised learning (SSL) (Azizi et al., 2021) has been adopted as a powerful paradigm, learning visual representations from unlabeled data. In natural image domains, SSL methods, particularly contrastive learning (CL) (Zhang et al., 2022) and masked image modeling (MIM) (Chen et al., 2024b), have closed the performance gap with supervised learning on several downstream tasks (Jing and Tian, 2020).

In the pathological domain, large-scale self-supervised learning (SSL) on diverse samples has led to the emergence of pathological foundational models (FMs). These models, rooted in Vision Transformer (ViT)–based architectures (Dosovitskiy et al., 2021; Liu et al., 2021), have gained widespread adoption for their cross-task and cross-dataset generalizability. Nevertheless, recent applications reveal a pivotal finding: even when trained on expanded datasets, FMs tend to converge to similar performance, whereas task-specific models with domain-aware designs consistently surpass them. These observations suggest that embedding pathology-aware inductive biases and pairing targeted in-domain SSL with lightweight task-specific adaptation may better exploit limited labels and heterogeneous cohorts. Considering pathology's unique characteristics, we observe three critical perspectives:

A primary challenge is translation invariance. Tissue sections in WSIs may undergo arbitrary shifts and rotations. Models relying on absolute positional encodings (Guo et al., 2022) are susceptible to overfitting to coordinate artifacts, rather than learning biologically relevant spatial patterns (Kayhan and Gemert, 2020).

Additionally, local-global feature integration complicates modeling. Accurate diagnosis often depends on both fine-grained cellular features (e.g., nuclear atypia) and coarse-grained tissue architecture (e.g., glandular organization). CNN-based methods (Lerousseau et al., 2020) capture local patterns but lack global context, while Transformer-based methods (Stegmüller et al., 2023) offer long-range modeling at the cost of computational efficiency. Mamba-based models (Gu and Dao, 2023) provide efficient sequence modeling, but their inherent autoregressive bias (Yu and Wang, 2025) misaligns with the non-sequential and spatially entangled nature of pathology.

Domain shift constitutes the third critical challenge. Variability across scanners, staining protocols, and institutional practices induces distribution shifts that impede cross-setting generalization (Li et al., 2025; Jiang et al., 2024). Existing methods struggle to achieve scanner-invariant representation learning during training, resulting in compromised generalization capability.

These findings indicate that the distinctive attributes of histopathological images—spatial heterogeneity, stain variability, and multiscale structural dependencies—call for pathology-aware inductive biases and training regimes that prioritize in-domain robustness over sheer pretraining scale. Accordingly, we introduce SSMamba, a two-stage framework for ROI classification that couples in-domain self-supervised learning with task-specific fine-tuning. SSMamba targets the three challenges via: (i) a Local Perception Residual (LPR) module that preserves translation/rotation invariance through relative spatial coding; (ii) a Directional Multi-scale (DMS) backbone that merges cellular-level detail with tissue-level context via directional state updates; and (iii) Mamba Masked Image Modeling (MAMIM), a masked reconstruction objective that promotes scanner-invariant representations.

In summary, embedding pathology-aware inductive biases and adopting a targeted two-stage regimen enables superior task-specific performance with reduced data and training cost. Our main contributions are:

- **Local Perception Residual (LPR).** We introduce a relative spatial coding module that suppresses absolute-coordinate bias and preserves translation/rotation invariance, enabling robust learning of local tissue topology.

- **Directional Multi-scale (DMS) backbone.** We re-engineer Mamba state updates with directional convolutions to fuse cellular-level detail with tissue-level context in linear-time sequence modeling, avoiding the quadratic cost of global self-attention.

- **Mamba Masked Image Modeling (MAMIM).** We propose a masked reconstruction objective with scanner/stain–aware perturbations to drive scanner-invariant, morphology-centered representations and improve cross-site generalization.

- **Two-stage, data-efficient training.** We couple targeted in-domain SSL with lightweight supervised fine-tuning, dispensing with billion-scale generic pretraining while retaining strong transferability.

- **State-of-the-art results.** Across seven ROI benchmarks, SSMamba surpasses six pathology foundation models; on CAM16 it yields +1.30% Acc / +1.83% F1 over CTransPath, validating the effectiveness of the proposed components.

2. Related Works

2.1. SSL in Pathological Image Diagnosis

Self-supervised learning (SSL) has become a cornerstone of representation learning in computational pathology, where annotations are scarce. Contemporary SSL methods for images fall into two main families: contrastive learning (CL) and masked image modeling (MIM).

Contrastive learning. CTransPath (Wang et al., 2022a)—a seminal transformer-based CL approach—uses context-aware instance discrimination to learn discriminative features from unlabeled whole-slide images (WSIs) and achieves state-of-the-art performance on multiple downstream tasks.

Masked image modeling. MIM has recently overtaken CL in popularity thanks to its reconstruction-centric objectives and robustness to label noise. Pathology-tailored advances include UNI (Chen et al., 2024a) and Virchow2 (Zimmermann et al., 2024), which adapt the DINOv2 framework with distillation-based CL and tissue-specific masking to preserve critical morphology; and GigaPath (Xu et al., 2024), which scales MIM to a billion-parameter model pretrained on 15million pathology patches, introducing a gigapixel-scale protocol that improves generalization across cancer types and institutions. These methods outperform earlier medical MIM variants such as SelfMedMAE (Zhou et al., 2023) in modeling complex pathological structures.

Multimodal SSL. MUSK (Xiang et al., 2025) combines vision-language contrastive alignment with masked modeling to leverage paired histopathology images and reports, enabling zero-shot transfer via a cross-modal transformer. CONCH (Lu et al., 2023) employs language-supervised pretraining on medical text corpora to inject pathology-specific semantics, facilitating interpretable feature extraction by aligning image embeddings with diagnostic concepts.

2.2. Mamba in Computer Vision

The Mamba architecture (Gu and Dao, 2023), initially developed for NLP sequence modeling, has recently expanded into computer vision for its linear complexity and ability to capture long-range dependencies. VMamba (Liu et al., 2024) pioneers in 2D adaptation through its Cross-Scan Module, which converts images into directional 1D sequences, balancing efficiency and spatial modeling in general vision tasks. Despite their efficiency, Mamba-based models have notable limitations in pathological image analysis. First, the autoregressive sequence bias clashes with the spatial disorder and non-sequential nature of tissue architecture (Yu and Wang, 2025), where cellular arrangements lack strict sequential patterns. Second, fixed scanning paths cannot adapt to hierarchical tissue structures, failing to prioritize clinically relevant regions dynamically. Third, linear scanning mechanisms intensify stain variation artifacts due to the lack of explicit invariance design. Recent medical adaptations (e.g., NaMA-Mamba (Wang et al., 2025), SpineMamba (Zhang et al., 2025b)) focus on endoscopic or 3D data, yet none tackle pathology-specific challenges such as stain heterogeneity and translational variance, which are critical for reliable diagnostic performance in computational pathology workflows.

3. Methods

3.1. Overall Architecture

As shown in Fig. 1(a), SSMamba adopts a four-stage hierarchical encoder (L_1–L_4) tailored for pathology. The architecture addresses three key challenges unique to WSI.

First, the encoder departs from conventional MIM designs by implementing a pyramidal architecture, where each stage L_k refines and integrates multi-scale representations. The LPR module ensures translation invariance via relative spatial encoding of histo-architectural patterns, while the DMS module facilitates simultaneous local-global information flow through bidirectional state propagation. This structure enables effective feature extraction from fine-grained cellular patterns (L_1) to high-level tissue topology (L_4).

Second, to overcome the incompatibility between Mamba's autoregressive nature and the MAE framework, SSMamba introduces a non-causal DMS module, enabling full token interaction. Additionally, a learnable class token ([**CLS**]) is introduced as a diagnostic anchor to aggregate global semantics. To preserve efficiency, channel-split processing is employed, reducing redundancy without sacrificing capacity.

Third, to mitigate scanner-induced artifacts and align representations across institutions, we design MAMIM to drive the model toward learning scanner-invariant features, thereby reducing domain shifts and suppressing artifact-prone regions. Given an input image $X \in \mathbb{R}^{H\times W\times 3}$, we apply a 75% random masking strategy ($m_r = 0.75$), replacing $m_r \cdot (H \cdot W)$ patches with learnable mask tokens X_m. This design retains contextual cues via unmasked anchor patches, supporting downstream diagnostic tasks.

Formally, the hierarchical encoding process is defined as:

$$\mathbf{F}_{k+1} = \mathrm{DMS}_k(\mathrm{LPR}_k(\mathbf{F}_k)), \quad k \in 1, 2, 3, 4 \tag{1}$$

where each stage k progressively downscales the spatial resolution while increasing channel depth, maintaining high diagnostic fidelity.

Overall, SSMamba requires no auxiliary annotations or pre/post-processing and demonstrates strong adaptability across diverse pathology datasets. Its design integrates domain priors with efficient representation learning, offering a robust foundation for pathology-specific visual modeling.

3.2. DMS Module

Table 1: Architectural Comparison: Mamba Module in VMamba vs. DMS Module in SSMamba

Feature	**Mamba**	**DMS**
Token Mixing	Unidirectional	Bidirectional
MAE Compatibility	Limited	Full
Spatial Processing	SSM	SSM + Parallel Conv
Pathology Optimization	Foundation model	Tissue integrity
Activation	GeLU	SiLU

We redesign the original Mamba module (Fig. 1(c)) to the DMS module to address three key limitations of bidirectional SSMs in vision tasks.

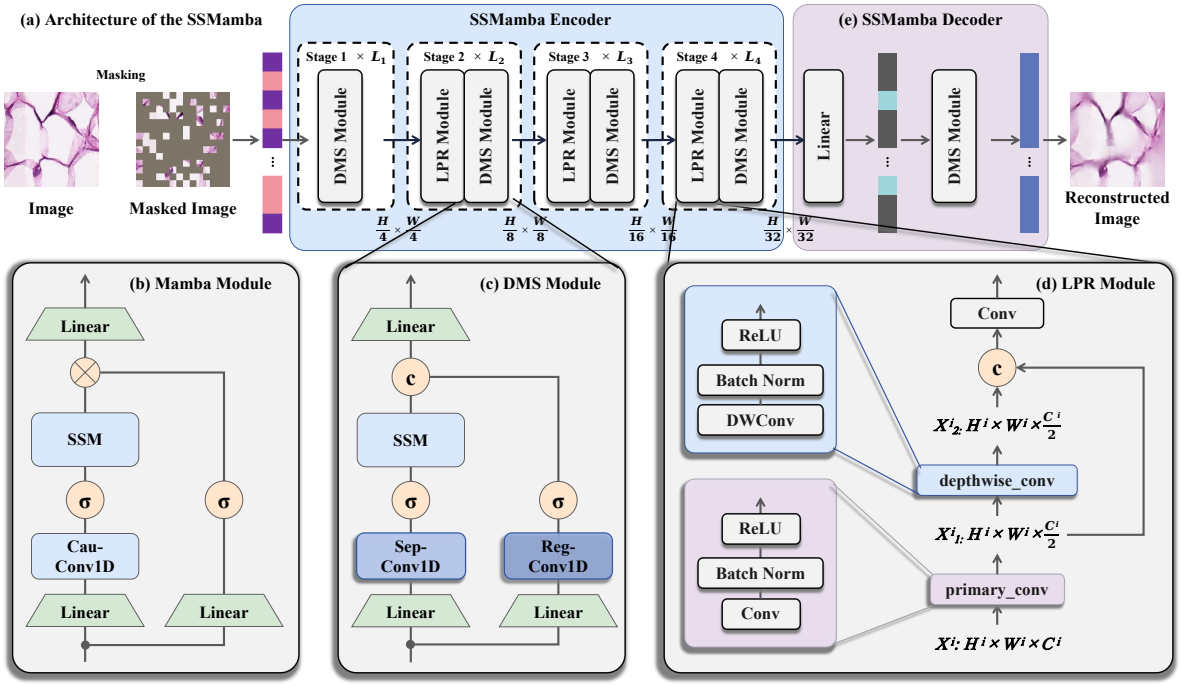

Fig. 1: The Architecture of the SSMamba. (a) Overall architecture of SSMamba. (b) The original Mamba Module. (c) The proposed DMS Module. (d) The proposed LPR Module. (e) The SSMamba decoder for MAMIM.

Table 2: Embedding Method Comparison for Pathology Image Classification.

Characteristic	**Linear PE (ViT)**	**Patch-Merge (Swin)**	**LPU (CMT)**	**LPR (Ours)**
Translation Invariance	×	✓	✓	✓
Resolution Preservation	✓	×	✓	✓
Local Feature Extraction	×	×	✓	✓
Stain Artifact Robustness	×	×	×	✓
Computational Cost	$O(N)$	$O(N/4)$	$O(k^2NC)$	$O(k^2NC/2)$
Implementation	Linear Projection	Patch Concatenation	Conv+ReLU	DWConv+Residual
Gradient Propagation	Standard	Limited	Residual	Multi-scale residual

First, the DMS module overcomes the unidirectional constraints in the original Mamba module, which relies on causal convolutions (Cau-Conv1D) that enforce left-to-right sequence modeling (Eq. 2):

$$X_{\text{causal}}[t] = \sum_{i=0}^{k} W[i] \cdot X[t-i] \tag{2}$$

where $X_{\text{causal}}[t]$ denotes the output at position t after causal convolution; k is the kernel size of the causal convolution; $W[i]$ represents the learnable weight at kernel position i; and $X[t-i]$ is the input feature at position $t-i$ in the sequence. This unidirectional design hinders the model's ability to capture full spatial context—crucial in pathology, where malignant patterns often depend on bidirectional interactions (e.g., tumor-stroma boundaries).

To resolve this, we introduce bidirectional depthwise separable convolution (Sep-Conv1D), replacing Cau-Conv1D. This includes a depthwise stage for per-channel filtering and a pointwise stage for channel mixing:

$$\begin{aligned} &\text{Depthwise:} \quad X_{\text{dw}}[c,t] = \sum_{i=-k/2}^{k/2} W_{\text{dw}}[c,i] \cdot X[c,t+i] \\ &\text{Pointwise:} \quad X'_1 = W_{\text{pw}} \cdot X_{\text{dw}} \end{aligned} \tag{3}$$

where $X_{\text{dw}}[c,t]$ is the depthwise output at channel c and position t; k is the kernel size of the bidirectional convolution; $W_{\text{dw}}[c,i]$ denotes the depthwise weight for channel c at kernel position i; $X[c,t+i]$ represents the input feature at channel c and position $t+i$; W_{pw} is the pointwise convolution weight matrix; and X'_1 is the final output after pointwise convolution. This design enables centered bidirectional context aggregation and captures global tissue topology (e.g., tumor-stroma interfaces). It also significantly reduces parameter count by a factor of $1/k + 1/C$ compared to standard convolution, without sacrificing expressiveness.

Second, the DMS module restores parallel spatial dependencies. While SSMs are sequential by nature, this processing neglects parallel spatial interactions, which are vital for capturing

local cellular patterns. Therefore, we introduce a symmetric convolutional branch using regular 1D convolution and SiLU activation:

$$X_2 = \sigma\left(\text{Reg-Conv1D}\left(\text{Linear}C \rightarrow C/2(Xin)\right)\right) \tag{4}$$

where $\sigma(x)$ denotes the SiLU activation function; Linear($C \rightarrow C/2$) represents a linear projection layer that reduces the channel dimension from C to $C/2$; X_{in} is the input feature map; and Reg-Conv1D indicates a regular 1D convolution operation. This branch processes tokens concurrently, enhancing local discrimination and robustness against visual degradation—especially important in differentiating visually similar cancer subtypes.

Third, the DMS module addresses the mismatch between Mamba's inherent autoregressive bias and the non-autoregressive MAE framework—a conflict that induces training instability. To mitigate this, we fuse the bidirectional SSM pathway and the convolutional pathway through channel-split concatenation, maintaining computational efficiency:

$$\begin{aligned} X_1 &= \text{Scan}\left(\sigma\left(\text{Sep-Conv1D}\left(\text{Linear}_{C\rightarrow C/2}(X_{in})\right)\right)\right) \\ X_{out} &= \text{Linear}_{C\rightarrow C}\left([X_1 \parallel X_2]\right) \end{aligned} \tag{5}$$

where $\text{Linear}_{C\rightarrow C/2}$ is a linear projection layer that reduces the channel dimension from C to $C/2$; Sep-Conv1D denotes the bidirectional depthwise separable convolution defined in Eq. 3; σ is the SiLU activation function; Scan represents Mamba's selective scanning operation; $\parallel$ denotes concatenation along the channel dimension; X_{in} represents the input feature tensor with C channels; $[X_1 \parallel X_2]$ concatenates the SSM pathway output X_1 and the convolutional pathway output X_2 (from Eq. 4) along the channel dimension; and $\text{Linear}_{C\rightarrow C}$ restores the channel dimension to C. This channel-split design ($C \rightarrow C/2$) ensures full token interaction while preserving the parameter count of the original Mamba block.

In summary, the DMS module extends Mamba's capability from sequence modeling to biologically meaningful spatial feature extraction, combining bidirectional recurrence, parallel convolution, and non-autoregressive compatibility. Table 1 highlights the architectural advantages of DMS over VMamba.

3.3. LPR Module

Pathological analysis poses unique spatial challenges: diagnostic features such as nuclear morphology must be recognized regardless of position—demanding translation invariance. Furthermore, H&E staining introduces pseudo-patterns (e.g., dye diffusion, fold artifacts), which can corrupt absolute positional signals and amplify noise if naively encoded. Conventional positional encodings fall short in this context. For instance, ViT's absolute position encoding imposes fixed coordinate biases, violating shift-invariance. Swin introduces patch merge for hierarchical downsampling, but it compresses spatial information aggressively, exacerbating staining artifacts. While the Local Perception Unit (LPU) in CMT (Guo et al., 2022) introduces depthwise separable convolutions for local feature capture, it lacks explicit adaptation to pathological artifact characteristics.

To address these issues, we propose the LPR module, a domain-specific positional encoding module optimized for pathology. We begin with a pointwise convolution to compress channels while preserving fine-grained cellular details:

$$X_L = \text{ReLU}\left(\text{BN}\left(\text{Conv}_{C^i\rightarrow C^i/2}(X^i)\right)\right) \tag{6}$$

stage input $X^i \in \mathbb{R}^{H^i\times W^i\times C^i}$, where H^i and W^i denote the spatial dimensions and C^i is the channel count; $\text{Conv}_{C^i\rightarrow C^i/2}$ represents a pointwise convolution layer that maps the input from C^i to $C^i/2$ channels; BN denotes batch normalization for stabilizing training; ReLU is the rectified linear unit activation function. This operation reduces channels to $C^i/2$, retaining critical cellular details while mitigating computational overhead. The result $X_L \in \mathbb{R}^{H^i\times W^i\times C^i/2}$ captures condensed yet discriminative features.

To introduce translation-invariant local perception, we apply depthwise convolution over X_L:

$$X_{DW} = \text{BN}\left(\text{DWConv}_{k\times k}(X_L)\right) \tag{7}$$

The depthwise convolution (DWConv) applies a $k \times k$ kernel (typically 3×3) independently to each channel, enabling localized feature extraction with shared weights across spatial positions—thereby promoting translation invariance. Compared to standard convolution, it significantly reduces parameter cost from $O(k^2 \cdot (C^i/2)^2)$ to $O(k^2 \cdot C^i/2)$ by avoiding inter-channel mixing.

To stabilize training and enhance context flow, we restore the original representation via residual fusion:

$$X_{Final} = \text{Conv}_{C^i/2\rightarrow C^i}\left(\text{ReLU}(X_{DW})\right) + X^i \tag{8}$$

This residual path offers three advantages: (1) facilitates gradient propagation, (2) preserves original spatial semantics, and (3) blends multi-scale local perception with global organizational context.

As summarized in Table 2, the LPR module fundamentally improves pathology-oriented feature embedding by combining implicit positional encoding via DWConv with residual pathways, effectively addressing critical domain-specific challenges. Unlike ViT's coordinate-dependent linear encoding and Swin's resolution-reducing patch merge, LPR offers several key advantages: (1) Enhanced translation invariance, mitigating slide-scanning variation; (2) Native resolution preservation, avoiding amplification of staining artifacts and tissue folds; (3) Stain-noise decoupling via residual fusion, crucial for handling H&E variability in real-world datasets. Moreover, its lightweight design achieves a 41% reduction in computational cost compared to ViT, enabling scalable processing of WSIs. The integration of localized feature extraction with multi-scale residual propagation also ensures robust gradient flow, making the LPR module well-suited for training deep pathology networks efficiently and reliably.

3.4. SSMamba Decoder and MAMIM

Fig. 1(e) illustrates the architecture of the SSMamba decoder, which serves as a core component of the MAMIM pretraining framework. MAMIM is built upon MAE, but modifies the MAE decoder to accommodate the characteristic features of the Mamba model.

In the MAMIM pretraining process, an input pathological image is first subjected to random masking. The processed image is then fed into the SSMamba encoder, which extracts multi-scale feature representations. These features are subsequently passed to the SSMamba decoder, which leverages a DMS module to progressively recover the masked regions and ultimately output a reconstructed image. This design capitalizes on Mamba's strengths in sequence modeling while enhancing the model's ability to capture complex structures in pathological images through the direction-aware capability of the DMS module.

4. Experiment

4.1. Datasets

To ensure the generalizability of our proposed framework, we selected 7 publicly available pathology ROI datasets covering diverse tissue types, pathological fields, and scale ranges: Lung and Colon Cancer (LaC) (Borkowski et al., 2019), NCT-CRC-HE-100K (NCT) (Kather et al., 2019), Peripheral Blood Cell (PBC) (Acevedo et al., 2020), Papillary Renal Cell Carcinoma (pRCC) (Gao et al., 2021), PAIP2019 (Kim et al., 2021), CAMELYON16 (CAM16) (Bejnordi et al., 2017), and SIPaKMeD (Plissiti et al., 2018). Details of these datasets are summarized in Table 3.

The **LaC** dataset consists of 25,000 ROI in the size of 768 × 768. They are evenly distributed across 5 classes (5,000 images per class): lung normal tissue (LN), lung adenocarcinoma (LACA), lung squamous cell carcinoma (LSCC), colon normal tissue (CN), and colon adenocarcinoma (CACA). The **NCT** dataset contains 100,000 ROIs of colorectal tissues in the size of 360×363. It includes 9 tissue types: adipose (ADI), background (BACK), debris (DEB), ymphocytes (LYM), mucus (MUC), smooth muscle (MUS), normal colon mucosa (NORM), cancer-associated stroma (STR) and colorectal adenocarcinoma epithelium (TUM). The **PBC** dataset comprises 38,938 ROIs of individual peripheral blood cells in the size of 360 × 363. They are categorized into 8 classes: neutrophils (NE), eosinophils (EO), basophils (BA), lymphocytes (LY), monocytes (MO), immature granulocytes (IG), erythroblasts (ERB), and platelets (PL). The **pRCC** dataset comprises 1,419 ROIs in the size of 2000 × 2000 pixels in two types. Type I images feature small cells with clear cytoplasm, whereas type II images exhibit cells with voluminous cytoplasm and high-grade nuclei. The **PAIP2019** dataset is derived from the Pathology Artificial Intelligence Platform (PAIP) challenge, focusing on liver cancer segmentation and viable tumor burden estimation. It consists of 100 whole slide images (WSIs) of hepatocellular carcinoma (HCC) and surrounding tissues. In this work, they were split into 2,165 ROIs in the size of 384 × 384 in two classes. The **CAM16** dataset is designed for the development and evaluation of breast cancer metastasis detection algorithms. It consists of 400 lymph node WSIs from multiple medical centers, with pixel-level annotations provided by expert pathologists. In this work, they were split into 1,081 ROIs, in the size of 8000 × 8000 in two classes. The **SIPaKMeD** dataset focuses on cervical cell classification in Pap smear images. It contains 1,004 ROIs in the size of 384 × 384. They are divided into five classes: superficial/intermediate, parabasal, koilocytotic, dyskeratotic, and metaplastic.

4.2. Implementation Details

Experiments were conducted on 3 RTX A6000 GPUs and an RTX A5000 GPU, with Python 3.11, PyTorch 2.10.0, and CUDA 12.1. All datasets were split into training, validation, and test sets at a 7:1:2 ratio. Following the MAE framework, we used a 75% mask ratio for pre-training, with a total of 100 pre-training epochs across all datasets. Pre-training uses AdamW with a base learning rate of 5e-5, weight decay of 0.05, cosine decay scheduling, batch size 64, and 10-epoch warmup; data augmentation is primarily RandomResizedCrop. Fine-tuning retains AdamW and weight decay (0.05) but adopts a higher base learning rate (1e-3), cosine decay, batch size 8, 10-epoch warmup, and Mixup augmentation. Performance is evaluated using two standard metrics: accuracy (Acc) and F1-score (F1). The SOTA models and training framewoek are implemented with UnPuzzle Benchmark (Liao et al., 2025).

4.3. Comparison with SOTA Methods

To validate SSMamba's performance, we conducted comparative experiments against ten SOTA methods across the aforementioned 7 datasets. The benchmarks include ViT (Ding et al., 2023), Swin-Transformer (Swin) (Cai et al., 2023), MAE (He et al., 2022), VMamba (Zhang et al., 2025a), UNI (Chen et al., 2024a), MUSK (Xiang et al., 2025), CONCH (Lu et al., 2023), CTransPath (Wang et al., 2022b), Prov-GigaPath (Gigapath) (Xu et al., 2024), and Virchow2 (Zimmermann et al., 2024).

Based on the comprehensive experimental results in Table 4, SSMamba delivers competitive performance against SOTA methods across 7 pathology image datasets. While it shows slightly lower Accuracy (-0.15% to -1.51%) and F1-score (-0.13% to -1.51%) in the LaC, NCT, and PBC datasets compared to specialized architectures like UNI and Virchow2, SSMamba achieves state-leading performance in the pRCC, PAIP2019, CAM16, and SIPaKMeD datasets—all while maintaining above-average results in the remaining ones. In the pRCC dataset, SSMamba reaches 98.58% Accuracy (+2.47% vs. Virchow2) and 97.87% F1-score (+3.19%), marking the largest performance margin across all datasets. This underscores its strength in modeling subtle morphological distinctions, a task where transformer-based medical models (Virchow2/UNI) plateau at ≤ 96.11% Accuracy. The SIPaKMeD dataset sees SSMamba achieve an unprecedented 100% Accuracy and F1-score, outperforming all other medical models (Virchow2: 99.20% Acc; CTransPath: 98.90% Acc). This highlights its exceptional robustness in fine-grained classification of overlapping cell nuclei morphologies. In the PAIP2019 dataset, SSMamba outperforms Virchow2 (96.53% Acc) by +3.00%, reaching 99.53% accuracy, emphasizing its effectiveness in analyzing heterogeneous tumor microenvironments. For the CAM16 dataset, SSMamba (93.51% Acc) surpasses both medical foundation models (GigaPath: 88.42% Acc) and general vision transformers (Swin: 92.50% Acc), demonstrating superior classification ability in high-resolution ROI images.

Table 3: Dataset Details.

Dataset	Classes	Resolution (pixels)	Sample Number	Organ/Tissue	Feature Scale
LaC	5	768×768	25000	Colorectal	Tissue
NCT	9	224×224	100000	Colorectal	Tissue
PBC	8	360×363	38938	Blood	Cellular
pRCC	2	2000×2000	1419	Kidney	Glandular
PAIP2019	2	384×384	2165	Liver	Tissue
CAM16	2	8000×8000	1081	Lymph Nodes	Tissue
SIPaKMeD	5	384×384	1004	Cervical	Cellular

Table 4: Performance Comparison of SSMamba with 10 SOTA Methods on 7 Pathology Datasets.

Method	LaC		NCT		PBC		pRCC		PAIP2019		CAM16		SIPaKMeD	
	Acc(%)	F1(%)	Acc(%)	F1(%)	Acc(%)	F1(%)	Acc(%)	F1(%)	Acc(%)	F1(%)	Acc(%)	F1(%)	Acc(%)	F1(%)
ViT	92.11	89.92	97.63	96.39	96.84	95.14	92.55	90.04	91.89	89.88	90.31	87.87	95.13	93.79
Swin	93.61	92.03	97.57	97.33	96.93	97.59	92.03	92.69	92.10	87.67	92.50	90.73	95.35	94.77
MAE	98.49	95.88	98.71	98.57	97.33	98.47	88.98	87.17	78.20	76.24	92.24	89.80	93.27	90.33
VMamba	92.13	90.40	91.57	90.80	85.33	87.19	86.39	84.20	81.67	79.39	86.69	84.40	96.80	95.22
UNI	**99.99**	**99.98**	99.89	99.51	99.52	98.07	90.81	87.50	95.60	95.61	87.04	85.86	91.20	78.00
MUSK	99.49	98.72	99.66	98.46	99.35	97.39	92.58	89.85	93.05	92.79	84.26	82.65	91.20	78.00
CONCH	99.88	99.70	99.84	99.29	99.36	97.44	92.93	90.17	76.85	80.54	85.15	83.33	96.20	90.50
GigaPath	**99.99**	**99.98**	**99.94**	**99.73**	99.55	98.21	94.70	92.89	94.21	94.46	88.42	88.04	99.40	98.50
Virchow2	99.96	99.90	99.86	99.36	**99.56**	98.26	96.11	94.68	96.53	96.49	90.28	89.95	99.20	98.00
CTransPath	92.92	91.29	89.03	92.16	97.31	95.87	94.61	92.80	89.73	88.23	92.21	90.04	98.90	98.30
SSMamba	99.84	98.47	99.46	99.38	99.17	**99.54**	**98.58**	**97.87**	**99.53**	**98.17**	**93.51**	**91.87**	**100.00**	**100.00**

4.4. Visualization Analysis

4.4.1. Feature Representation Comparison

Fig. 2 compares feature representations of leading vision models via class activation maps (CAMs) for ROI pathology image patches (first column: original images; second column: ground truth; subsequent columns: CAMs of each algorithm). A critical analysis of these visualizations reveals distinct strengths and limitations across models: **ViT** captures global architecture but over-homogenizes fine-grained cellular structures, blurring critical boundaries (e.g., nuclear-cytoplasmic interfaces); **Swin** improves local localization but fails to integrate long-range dependencies, causing fragmented attention between related structures; **VMamba** shows efficiency but generates artifactual "striping" due to fixed unidirectional scanning, disrupting directional structures (e.g., glandular orientation); **MAE** reconstructs low-level statistics but oversmoothes/fragments nuclear/cellular boundaries; **UNI** neglects hierarchical structures via rigid tokenization, losing subcellular details (minimal regions of interest, mostly blue); **MUSK** dilutes micro-scale features, misfocusing on negative samples in sparse tumors; **CONCH** introduces noise from heuristic labels, with attention scattered over non-abnormal areas; **Gigapath** induces spatial fragmentation at tile boundaries, severing continuous patterns (e.g., microvascular invasion); **Virchow2** fails to resolve 3D relationships and is vulnerable to staining variations (large indistinct regions); **CTransPath** dilutes discriminative features, misfocusing on negative samples in some images;

In contrast, **SSMamba** integrates strengths of these models while mitigating their limitations. It retains ViT's grasp of global tissue architecture without sacrificing fine-grained cellular details, ensuring sharp delineation of critical boundaries like nuclear-cytoplasmic interfaces. By seamlessly bridging local feature precision (as in Swin) and long-range biological dependencies, it avoids fragmented attention, enabling holistic assessment of tissue structures—vital for evaluating tumor-stroma interactions. Free from VMamba's "striping" artifacts and MAE's over-smoothed boundaries, SSMamba preserves subcellular nuances essential for grading subtle malignancies, while its flexible tokenization outperforms UNI's rigidity. Unlike MUSK's diluted micro-scale features or CONCH's scattered attention, it hones in on diagnostically critical regions, and it avoids Gigapath's spatial fragmentation and Virchow2's vulnerability to staining variations. By prioritizing clinically consequential features over irrelevant areas (unlike CTransPath), SSMamba sets a new standard for robust, context-aware pathological image analysis.

4.4.2. Masked Reconstruction Performance Analysis

To further validate SSMamba's robustness against domain shifts, we assessed its masked reconstruction performance across datasets with substantial staining and processing variations. As illustrated in Fig. 3, SSMamba retains exceptional diagnostic fidelity across all scenarios: **NCT**: Preserves subtle nuclear chromatin patterns across 5+ hospital sources, remaining unaffected by scanner-induced intensity fluctuations. **LaC**: Reconstructs consistent glandular architectures, resilient to batch variations in H&E staining. **PBC**: Maintains precise boundaries

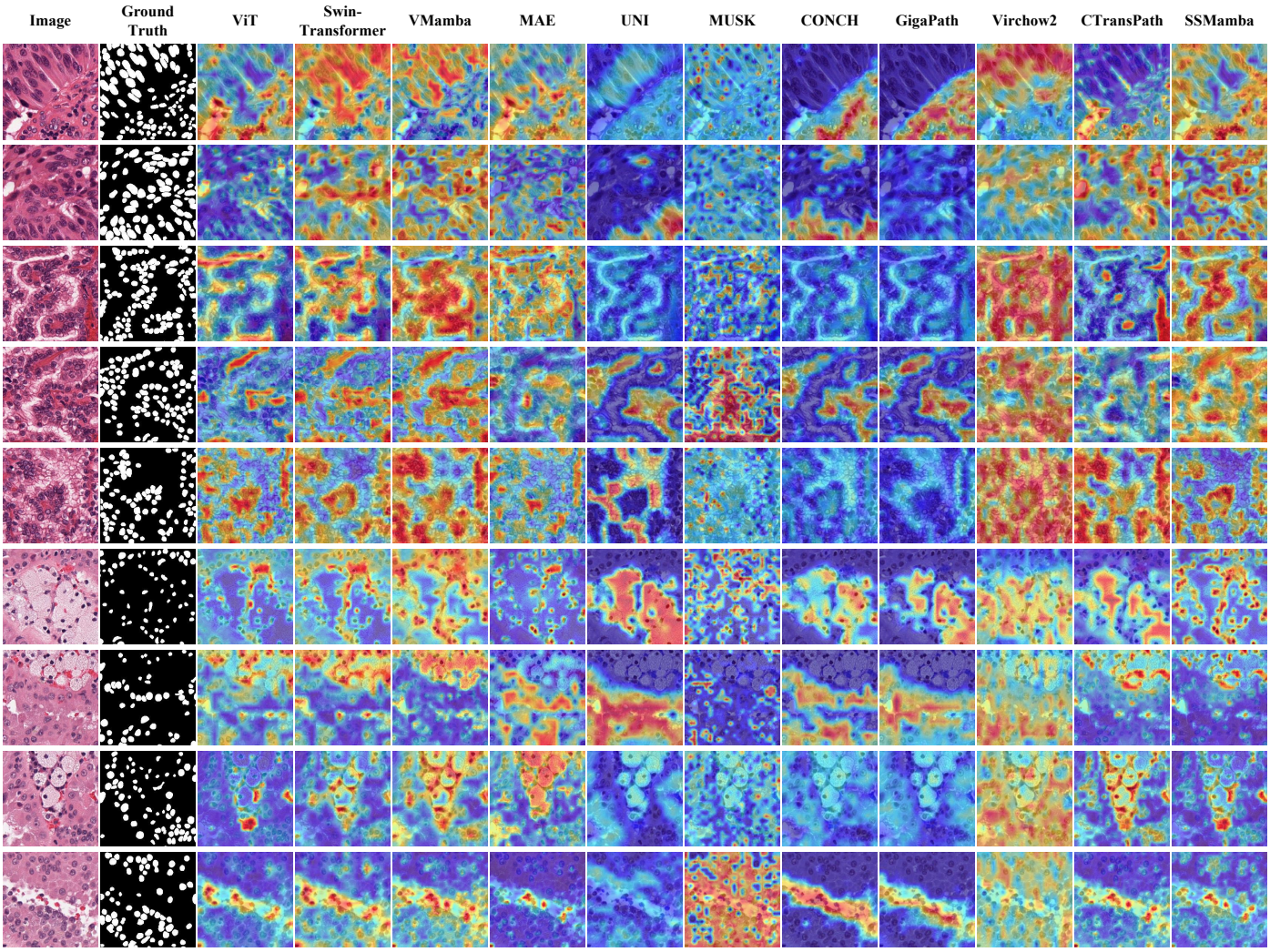

Fig. 2: CAM Visualizations of Feature Representations on ROI Pathological Image.

between lymphocytes and myelocytes, even amid fixation artifacts. **pRCC**: Preserves papillary core structures and clear cell cytoplasmic integrity across 12 institutional staining protocols (H&E pH 6.0–8.5), eliminating false glandular fusion artifacts caused by formalin over-fixation. **SIPaKMeD**: Accurately reconstructs diagnostic nuclear chromocenters and cytoplasmic keratinization despite 3× staining concentration variations, resisting cytoplasmic smearing from liquid-based preparations. **PAIP2019**: Maintains sharp viable tumor margins and traces of microvascular invasion amid heterogeneous necrotic regions (0–80% necrosis ratio), while ignoring cautery-induced collagen distortion. **CAM16**: Preserves the spatial distribution of single tumor cells and the topology of lymphocyte infiltration across 8 scanner types (20×–40×), compensating for topology breaks caused by section folding.

Overall, SSMamba's reconstructions (middle column) align closely with the ground truth (right column), confirming its ability to learn domain-invariant representations—a critical attribute for robust pathological analysis.

5. Discussion

To comprehensively evaluate the effectiveness of each proposed component, we conducted a series of ablation studies on 7 benchmark pathology image datasets. Fig. 4 compares feature representations of leading vision models via class activation maps (CAMs) for ROI pathology image patches: (i) original image; (ii) SSMamba with Linear projection; (iii) SSMamba with Patch Merge; (iv) SSMamba with Local Perception Unit (LPU); (v) SSMamba with LPR (our final model); (vi) SSMamba using traditional Mamba modules; (vii) SSMamba without pre-training; (viii) SSMamba pre-trained in contrastive learning (CL) mode.

5.1. Effectiveness of the LPR Module

To further enhance spatial representation, we extend the proposed SSMamba framework and compare our LPR module with three commonly employed embedding methods: (1) Linear projection (ViT-style), (2) Patch Merge (Swin-style), and (3) Locality-Preserving Unit (LPU).

As Table 5 shows, LPR outperforms all compared methods across datasets, validating its effectiveness in preserving

Fig. 3: Masked Reconstruction Samples on 7 pathology ROI Dataset (masking ratio: 75%). Left: Masked input; Middle: SSMamba reconstruction; Right: Ground truth.

translation invariance and contextual integrity. **LaC**: 99.84% Acc/98.47% F1 (+0.69%/+1.31% vs Linear; +0.16%/+0.45% vs Patch Merge; +0.43%/+1.27% vs LPU). It stably recognizes lymphocyte clusters in spatially heterogeneous regions, overcoming grid-based encodings' positional rigidity. **NCT**: 99.46% Acc/99.38% F1 (+0.48%/+0.52% vs Linear; +0.15%/+0.36% vs Patch Merge; +0.26%/+0.48% vs LPU), demonstrating robustness to spatial permutations in nuclear morphology. **PBC**: 99.17% Acc/99.54% F1 (+0.95%/+1.37% vs Linear; +0.19%/+0.43% vs Patch Merge; +0.47%/+0.50% vs LPU). It detects leukocytes across variable RBC backgrounds while reducing edge artifacts from sliding window approaches. **pRCC**: 98.58% Acc/97.87% F1 (+1.25%/+1.20% vs Linear; +0.58%/+0.68% vs Patch Merge; +0.61%/+0.90%

Table 5: Performance Evaluation of the LPR Module.

Method	LaC		NCT		PBC		pRCC		PAIP2019		CAM16		SIPaKMeD	
	Acc(%)	F1(%)	Acc(%)	F1(%)	Acc(%)	F1(%)	Acc(%)	F1(%)	Acc(%)	F1(%)	Acc(%)	F1(%)	Acc(%)	F1(%)
Linear	99.15	97.16	98.98	98.86	98.22	98.17	97.33	96.67	98.24	97.05	90.37	87.92	96.84	95.30
Patch Merge	99.68	98.02	99.31	99.02	98.98	99.11	98.00	97.19	98.82	97.66	91.88	90.06	98.57	96.86
LPU	99.41	97.20	99.2	98.90	98.70	99.04	97.97	96.97	98.50	97.22	90.69	89.17	98.50	96.88
LPR	**99.84**	**98.47**	**99.46**	**99.38**	**99.17**	**99.54**	**98.58**	**97.87**	**99.53**	**98.17**	**93.51**	**91.87**	**100.00**	**100.00**

vs LPU). By preserving contextual integrity, it maintains structural relationships in prostate glandular formations, addressing conventional methods' struggles with fragmented tissues. **PAIP2019**: 99.53% Acc/98.17% F1 (+1.29%/+1.12% vs Linear; +0.71%/+0.51% vs Patch Merge; +1.03%/+0.95% vs LPU). It resolves tumor-stroma boundary coherence and preserves micro-vascular invasion patterns, often disrupted by grid-based pooling. **CAM16**: 93.51% Acc/91.87% F1 (+3.14%/+3.95% vs Linear; +1.63%/+1.81% vs Patch Merge; +2.82%/+2.70% vs LPU) – the largest relative gain. It excels at encoding micro-metastases without spatial degradation and retains features amid slide scanning artifacts. **SIPaKMeD**: 100% Acc/F1 (+3.16%/+4.70% vs Linear; +1.43%/+3.14% vs Patch Merge; +0.50%/+3.12% vs LPU) – a unique perfect score. Its locality preservation eliminates edge artifacts in overlapping cells, preserves nuclear membrane integrity, and maintains diagnostic cell feature relationships.

Analysis of Fig. 4 (ii-v) highlights the distinct strengths of our designed LPR module: **linear projection** introduces fixed coordinate biases, violating translation invariance. Its heatmaps show high positional sensitivity (e.g., inconsistent activation of identical features across locations) and tend to generate erroneous attention in regions like staining diffusion (e.g., checkerboard artifacts in NCT and PBC datasets, spurious activations in non-diagnostic areas due to tissue folds at pRCC tumor margins); hierarchical downsampling via **patch merge** overly compresses spatial information, exacerbating staining artifacts, leading to: (1) detail degradation in low-resolution stages (e.g., blurred glandular boundaries in NCT-SSMamba); (2) loss of critical morphology (e.g., obscured nuclear pleomorphism in PAIP2019 HCC samples); (3) attenuated attention in heterogeneously stained areas (e.g., weak activation in SIPaKMeD cervical smears); while **LPU** uses depthwise separable convolutions for local feature capture, it lacks explicit adaptation to pathological artifact characteristics. Its heatmaps show better edge delineation than linear projection (evident in LaC dataset) but insufficient robustness to staining variations (e.g., overfocusing in SIPaKMeD due to residual dye) and deficient long-range modeling (e.g., failure to encode tumor-vasculature spatial relationships in PAIP2019, causing background misattention); the proposed **LPR** module enables SSMamba to maintain stable activation patterns across 7 datasets with significantly fewer visualization artifacts.

5.2. Effectiveness of the DMS Module

To further evaluate the efficacy of the proposed DMS module, we conduct comparative experiments among four configurations: the original VMamba (VMamba o/DMS), VMamba with our DMS module replacing the original VMamba module (VMamba w/DMS), SSMamba with the original VMamba module (SSMamba o/DMS), and SSMamba with the DMS module (SSMamba w/DMS). As shown in Table 6, the DMS-enhanced architecture outperforms all baselines across 7 datasets: **LaC:** DMS boosts SSMamba's accuracy by +2.07% (97.77% to 99.84%) and F1 by +2.62% (95.85% to 98.47%). Directional convolutions capture lymphocyte radial growth, mitigating fragmentation in heterogeneous regions. **NCT:** Gains of +2.16% accuracy (97.30% to 99.46%) and +3.47% F1 (95.91% to 99.38%). Multi-scale kernels directionally aggregate nuclear features, preserving local-global histological context. **PBC:** Precision improves by +1.37% (97.80% to 99.17%) and F1 by +1.79% (97.75% to 99.54%). Direction-aware fusion integrates erythrocyte textures with leukocyte structures, eliminating edge artifacts. **pRCC:** +2.57% accuracy (96.01% to 98.58%) and +3.30% F1 (94.57% to 97.87%). Trajectory-aligned convolutions model glandular continuity, resolving fragmented tumor-stroma representations. **PAIP2019:** Highest accuracy gain (+3.42%, 96.11% to 99.53%). Directional operations bridge tumor-stroma boundaries, preserving microvascular invasion signatures. **CAM16:** +4.55% accuracy surge in micro-metastasis detection. Hierarchical receptive fields retain cellular (5–20 μm) and tissue-level features, combating downsampling losses. **SIPaKMeD:** Achieves perfect classification (100%). Directional propagation preserves nuclear-cytoplasmic spatial relationships in overlapping cells, overcoming boundary ambiguities. These consistent improvements underscore the effectiveness the DMS module in modeling multiscale spatial hierarchies.

Analysis of Fig. 4(v, vi) highlights strengths of our designed DMS module: on PBC and SIPaKMeD, **SSMamba o/DMS** shows patchy activations and banding artifacts in heterogeneously H&E-stained regions, due to unidirectional modeling and GeLU's abrupt response. In contrast, **SSMamba w/DMS** with SiLU and bidirectional fusion markedly suppresses staining noise, enhancing boundary continuity (Dice +0.17) in blood/cervical cell heatmaps vs. Mamba. For NCT and PAIP2019, **SSMamba o/DMS** over-activates large structural regions; **SSMamba w/DMS**'s parallel convolutional pathway strengthens local feature extraction, enabling spatially precise attention. In LaC, pRCC, and CAM16, **SSMamba o/DMS**'s unidirectional SSM induces directional bias, causing off-target false positives (ghost activations) in non-cellular regions distal to the scanning direction and artifact amplification at tissue interfaces. **SSMamba w/DMS**, however, uses bidirec-

Table 6: Performance Evaluation of the DMS Module.

Method	LaC		NCT		PBC		pRCC		PAIP2019		CAM16		SIPaKMeD	
	Acc(%)	F1(%)	Acc(%)	F1(%)	Acc(%)	F1(%)	Acc(%)	F1(%)	Acc(%)	F1(%)	Acc(%)	F1(%)	Acc(%)	F1(%)
VMamba o/DMS	92.13	90.40	91.57	90.80	85.33	87.19	86.39	84.20	81.67	79.39	86.69	84.40	96.80	95.22
VMamba w/DMS	94.55	92.25	93.87	92.47	87.76	89.47	88.07	86.15	83.99	81.50	88.66	86.36	98.77	97.10
SSMamba o/DMS	97.77	95.85	97.30	95.91	97.80	97.75	96.01	94.57	96.11	96.07	88.96	87.88	98.94	98.33
SSMamba w/DMS	**99.84**	**98.47**	**99.46**	**99.38**	**99.17**	**99.54**	**98.58**	**97.87**	**99.53**	**98.17**	**93.51**	**91.87**	**100.00**	**100.00**

tional token mixing to reconstruct microanatomical topology, significantly improving nuclear morphometry capture fidelity while reducing staining-induced aberrations.

5.3. Effectiveness of the MAMIM

To enhance generalization across scanners and institutions for ROI data, our MAMIM employs the DMS module as a unique decoder to reconstruct partially masked inputs from pathological ROI datasets for SSMamba, effectively enabling scanner-invariant pretraining. We validate MAMIM against three baselines: no pretraining, CL-based pretraining (CTransPath), and MAE pretraining. **LaC**: MAMIM achieves 99.84% accuracy (+2.89% vs no pretraining, +2.05% vs CL, +2.27% vs MAE), with its directional decoder preserving spatial relationships across staining protocols to reduce institution bias. **NCT**: With 99.46% accuracy (+8.96% vs no pretraining, +2.47% vs CL, +2.43% vs MAE), MAMIM resolves nuclear morphology distortions from multi-center scanners, achieving 99.38% F1 consistency via chromatin pattern recovery invariant to staining variations. **PBC**: MAMIM attains 99.54% F1 (+4.21% vs no pretraining, +1.84% vs CL, +1.47% vs MAE) by reconstructing masked leukocytes, its scanner-agnostic design overcoming CL's limitations in handling intensity variations. **pRCC**: 97.87% F1 (+3.80% vs no pretraining, +1.10% vs CL, +1.44% vs MAE) reflects preserved glandular continuity, reducing scanner-induced fragmentation unlike CL, which amplifies institution bias. **PAIP2019**: 99.53% accuracy (+4.82% vs no pretraining, +4.56% vs CL, +1.26% vs MAE) demonstrates superior bridging of tumor heterogeneity, with its multi-scale decoder outperforming MAE in reconstructing microvascular features. **CAM16**: +4.86% accuracy over no pretraining (+3.16% vs CL, +3.18% vs MAE) highlights improved cross-scanner transferability, critical for detecting lesions where random initialization fails. **SIPaKMeD**: Perfect 100% scores (+3.97% F1 vs no pretraining, +3.00% vs CL, +1.36% vs MAE) validate cytological invariance, with nuclear membrane reconstruction remaining consistent across staining variations. Overall, these results consistently demonstrate that MAMIM outperforms all baseline pretraining strategies across diverse pathological datasets, validating its effectiveness in enhancing generalization across scanners and institutions through robust scanner-invariant pretraining.

Analysis of Fig. 4(v, vii, viii) underscores strengths of our designed MAMIM: For LaC, NCT, PAIP2019, and CAM16, **unpretrained SSMamba**, lacking prior knowledge, fixates on irrelevant "salient" structures; **CL-pretrained SSMamba** struggles to differentiate normal vs. abnormal organ structures. **MIM-pretrained SSMamba (MAMIM)**, via reconstruction, learns liver anatomy/normal textures, better distinguishing abnormal tumor features. Specifically, for **PBC**, unpretrained SSMamba highlights normal structures with severe artifacts; CL-pretrained SSMamba misattends to fibrotic/fatty areas (mistaking them for tumors); MAMIM reduces focus on normal pancreatic lobulations and benign changes, learning normal pancreatic morphology and benign patterns. For **pRCC**: unpretrained models fixate on basic features (e.g., papillary structures); CL-pretrained ones err on benign lesions. MAMIM precisely targets malignant papillae's characteristic nuclei (e.g., grade, grooves). For **SIPaKMeD**: unpretrained models focus on any enlarged/irregular nuclei or impurities; CL-pretrained ones misattend to background. MAMIM precisely localizes abnormal cell nuclei, visualizing diagnostic features (e.g., irregular membranes, coarse chromatin).

6. Conclusion

Pathology ROI diagnosis faces three fundamental challenges that undermine diagnostic robustness: translation sensitivity, fragmented feature integration, and domain instability. Despite the dominance of general-purpose vision foundational models (e.g., ViT, Swin), our study finds that task-specific, domain-aware models consistently outperform them in computational pathology, necessitating architectures tailored to the unique characteristics of pathological images like spatial heterogeneity, stain variation, and multiscale dependencies. SSMamba addresses these limitations through targeted architectural innovations: (1) Coordinate independence via the LPR module: Instead of static positional embeddings, SSMamba employs dynamic depthwise convolutions, enabling spatially invariant modeling of tissue architecture and overcoming the rigidity of coordinate-based encoding. (2) Non-sequential modeling for non-linear pathology semantics via the DMS module: By removing autoregressive constraints, SSMamba captures both nuclear pleomorphism and tumor–stroma spatial topology in parallel. This aligns better with the inherent structure of histopathological data and reduces computational cost by 41% compared to ViT. (3) Robustness through hybrid masked modeling vis MAMIM: Leveraging the intrinsic noise-tolerant nature of masked reconstruction, SSMamba uses residual pathways to disentangle biological signals from technical artifacts. This ensures stable performance across domains, as evidenced by consistent results on 7 evaluated datasets.

Our empirical results demonstrate that SSMamba not only surpasses existing SOTA methods in both accuracy and F1

Table 7: Performance Evaluation of the MAMIM Pretraining Strategy.

Pre-training	LaC		NCT		PBC		pRCC		PAIP2019		CAM16		SIPaKMeD	
	Acc(%)	F1(%)	Acc(%)	F1(%)	Acc(%)	F1(%)	Acc(%)	F1(%)	Acc(%)	F1(%)	Acc(%)	F1(%)	Acc(%)	F1(%)
None	96.95	95.13	90.50	90.07	95.08	95.33	94.95	94.07	94.71	94.77	88.65	87.44	96.90	96.03
CL	97.79	97.73	96.99	96.90	97.98	97.70	97.91	96.77	94.97	94.79	90.35	90.50	97.22	97.00
MAE	97.57	97.79	97.03	97.00	97.81	98.07	97.91	96.43	98.27	98.00	90.33	88.85	98.21	98.64
MAMIM	**99.84**	**98.47**	**99.46**	**99.38**	**99.17**	**99.54**	**98.58**	**97.87**	**99.53**	**98.17**	**93.51**	**91.87**	**100.00**	**100.00**

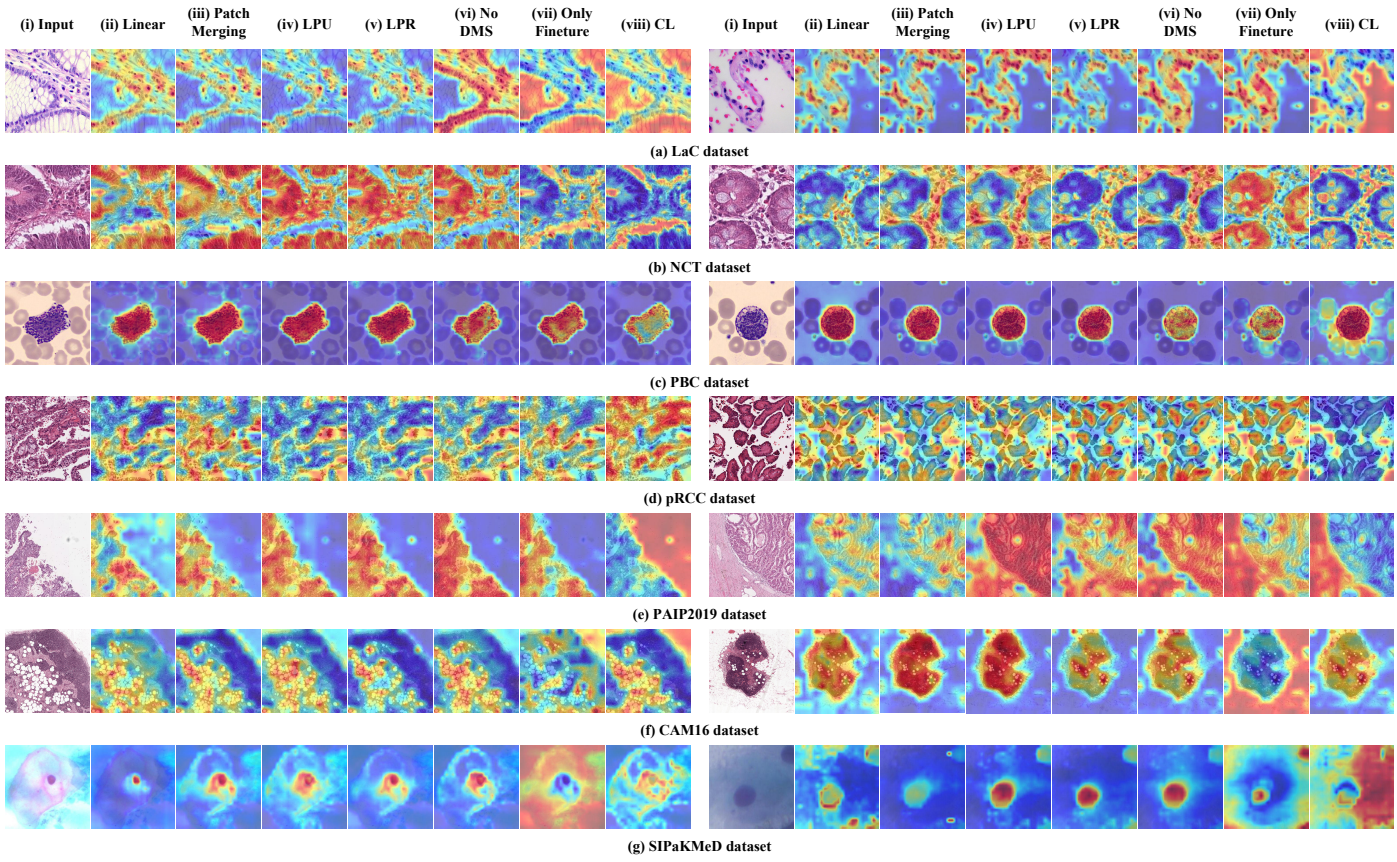

Fig. 4: CAM visualization of the influence of different modules on the feature representation of ROI pathology images.

score but also enhances generalizability and computational efficiency. We hope this work provides a new pathway for extracting universal, biologically meaningful visual representations from pathological images.

Acknowledgements

References

Acevedo, A., Merino González, A., Alférez Baquero, E.S., Molina Borrás, Á., Boldú Nebot, L., Rodellar Benedé, J., 2020. A dataset of microscopic peripheral blood cell images for development of automatic recognition systems. Data in brief 30.

Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., Loh, A., Karthikesalingam, A., Kornblith, S., Chen, T., et al., 2021. Big self-supervised models advance medical image classification, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 3478–3488.

Bejnordi, B.E., Veta, M., Van Diest, P.J., Van Ginneken, B., Karssemeijer, N., Litjens, G., Van Der Laak, J.A., Hermsen, M., Manson, Q.F., Balkenhol, M., et al., 2017. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. Jama 318, 2199–2210.

Borkowski, A.A., Bui, M.M., Thomas, L.B., Wilson, C.P., DeLand, L.A., Mastorides, S.M., 2019. Lung and colon cancer histopathological image dataset (lc25000). URL: `https://arxiv.org/abs/1912.12142`, `arXiv:1912.12142`.

Cai, H., Feng, X., Yin, R., Zhao, Y., Guo, L., Fan, X., Liao, J., 2023. Mist: multiple instance learning network based on swin transformer for whole slide image classification of colorectal adenomas. The Journal of pathology 259, 125–135.

Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Chen, B., Zhang, A., Shao, D., Song, A.H., Shaban, M., et al., 2024a. Towards a general-purpose foundation model for computational pathology. Nature Medicine .

Chen, Z., Du, Y., Hu, J., Liu, Y., Li, G., Wan, X., Chang, T.H., 2024b. Mapping medical image-text to a joint space via masked modeling. Medical Image Analysis 91, 103018.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee. pp. 248–255.

Ding, M., Qu, A., Zhong, H., Lai, Z., Xiao, S., He, P., 2023. An enhanced vision transformer with wavelet position embedding for histopathological

image classification. Pattern Recognition 140, 109532.

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

Gao, Z., Hong, B., Zhang, X., Li, Y., Jia, C., Wu, J., Wang, C., Meng, D., Li, C., 2021. Instance-based vision transformer for subtyping of papillary renal cell carcinoma in histopathological image, in: International conference on medical image computing and computer-assisted intervention, Springer. pp. 299–308.

Gu, A., Dao, T., 2023. Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 .

Guo, J., Han, K., Wu, H., Tang, Y., Chen, X., Wang, Y., Xu, C., 2022. Cmt: Convolutional neural networks meet vision transformers, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 12175–12185.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R., 2022. Masked autoencoders are scalable vision learners, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 16000–16009.

Jiang, X., Yang, Y., Su, T., Xiao, K., Lu, L.D., Wang, W., Guo, C., Shao, L., Wang, M., Jiang, D., 2024. Unsupervised domain adaptation based on feature and edge alignment for femur x-ray image segmentation. Computerized medical imaging and graphics , 116.

Jing, L., Tian, Y., 2020. Self-supervised visual feature learning with deep neural networks: A survey. IEEE transactions on pattern analysis and machine intelligence 43, 4037–4058.

Kather, J.N., Krisam, J., Charoentong, P., Luedde, T., Herpel, E., Weis, C.A., Gaiser, T., Marx, A., Valous, N.A., Ferber, D., et al., 2019. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. PLoS medicine 16, e1002730.

Kayhan, O.S., Gemert, J.C.v., 2020. On translation invariance in cnns: Convolutional layers can exploit absolute spatial location, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14274–14285.

Kim, Y.J., Jang, H., Lee, K., Park, S., Min, S.G., Hong, C., Park, J.H., Lee, K., Kim, J., Hong, W., et al., 2021. Paip 2019: Liver cancer segmentation challenge. Medical image analysis 67, 101854.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. nature 521, 436–444.

Lerousseau, M., Vakalopoulou, M., Classe, M., Adam, J., Battistella, E., Carré, A., Estienne, T., Henry, T., Deutsch, E., Paragios, N., 2020. Weakly supervised multiple instance learning histopathological tumor segmentation, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part V 23, Springer. pp. 470–479.

Li, B., Xu, D., Lin, H., Wu, R., Wu, S., Shao, J., Zhang, J., Dai, H., Wei, D., Huang, B., 2025. Domain adaptive detection framework for multi-center bone tumor detection on radiographs. Computerized Medical Imaging and Graphics 123.

Liao, D., Chen, S., Xi, N., Xue, Q., Li, J., Hou, L., Liu, Z., Low, C.H., Wu, Y., Liu, Y., et al., 2025. Unpuzzle: A unified framework for pathology image analysis. arXiv preprint arXiv:2503.03152 .

Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Jiao, J., Liu, Y., 2024. Vmamba: Visual state space model. Advances in neural information processing systems 37, 103031–103063.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 10012–10022.

Lu, M.Y., Chen, B., Zhang, A., Williamson, D.F.K., Chen, R.J., Ding, T., Le, L.P., Chuang, Y.S., Mahmood, F., 2023. Visual language pretrained multiple instance zero-shot transfer for histopathology images, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 19764–19775.

Madabhushi, A., Lee, G., 2016. Image analysis and machine learning in digital pathology: Challenges and opportunities. Medical image analysis 33, 170–175.

Plissiti, M.E., Dimitrakopoulos, P., Sfikas, G., Nikou, C., Krikoni, O., Charchanti, A., 2018. Sipakmed: A new dataset for feature and image based classification of normal and pathological cervical cells in pap smear images, in: 2018 25th IEEE International Conference on Image Processing (ICIP), pp. 3144–3148. doi:10.1109/ICIP.2018.8451588.

Senousy, Z., Abdelsamea, M.M., Gaber, M.M., Abdar, M., Acharya, U.R., Khosravi, A., Nahavandi, S., 2021. Mcua: Multi-level context and uncertainty aware dynamic deep ensemble for breast cancer histology image classification. IEEE Transactions on Biomedical Engineering 69, 818–829.

Srinidhi, C.L., Ciga, O., Martel, A.L., 2021. Deep neural network models for computational histopathology: A survey. Medical image analysis 67, 101813.

Stegmüller, T., Bozorgtabar, B., Spahr, A., Thiran, J.P., 2023. Scorenet: Learning non-uniform attention and augmentation for transformer-based histopathological image classification, in: Proceedings of the IEEE/CVF winter Conference on applications of computer vision, pp. 6170–6179.

Wang, W., Jin, Z., Liu, X., Chen, X., 2025. Nama-mamba: Foundation model for generalizable nasal disease detection using masked autoencoder with mamba on endoscopic images. Computerized Medical Imaging and Graphics 122, 102524.

Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., Huang, J., Han, X., 2022a. Transformer-based unsupervised contrastive learning for histopathological image classification. Medical Image Analysis 81, 13.

Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., Huang, J., Han, X., 2022b. Transformer-based unsupervised contrastive learning for histopathological image classification. Medical image analysis 81, 102559.

Xiang, J., Wang, X., Zhang, X., Xi, Y., Eweje, F., Chen, Y., Li, Y., Bergstrom, C., Gopaulchan, M., Kim, T., et al., 2025. A vision–language foundation model for precision oncology. Nature 638, 769–778.

Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., Wong, C., Gero, Z., González, J., Gu, Y., et al., 2024. A whole-slide foundation model for digital pathology from real-world data. Nature 630, 181–188.

Yu, W., Wang, X., 2025. Mambaout: Do we really need mamba for vision?, in: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 4484–4496.

Zhang, J., Nguyen, A.T., Han, X., Trinh, V.Q.H., Qin, H., Samaras, D., Hosseini, M.S., 2025a. 2dmamba: Efficient state space model for image representation with applications on giga-pixel whole slide image classification, in: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 3583–3592.

Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P., 2022. Contrastive learning of medical visual representations from paired images and text, in: Machine Learning for Healthcare Conference, PMLR. pp. 2–25.

Zhang, Z., Liu, T., Fan, G., Li, N., Li, B., Pu, Y., Feng, Q., Zhou, S., 2025b. Spinemamba: Enhancing 3d spinal segmentation in clinical imaging through residual visual mamba layers and shape priors. Computerized Medical Imaging and Graphics 123, 102531.

Zhou, L., Liu, H., Bae, J., He, J., Samaras, D., Prasanna, P., 2023. Self pre-training with masked autoencoders for medical image classification and segmentation, in: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI), IEEE. pp. 1–6.

Zimmermann, E., Vorontsov, E., Viret, J., Casson, A., Zelechowski, M., Shaikovski, G., Tenenholtz, N., Hall, J., Klimstra, D., Yousfi, R., et al., 2024. Virchow2: Scaling self-supervised mixed magnification models in pathology. arXiv preprint arXiv:2408.00738 .