MERIT: Multi-Scale Mamba with Enhanced Information Retention for Treatment Response Prediction from Whole Slide Images

Taiyuan Hu[1,2,†], Haijing Luan[1,2,†], Jifang Hu[1,2], Han Li[3], Yue Pei[1,2], Kaixing Yang[2], Ruilin Li[1], Beifang Niu[1], Xuebin Chi[1], Jinrong Jiang[1,2,⋆], and Rui Yan[4,5,⋆]

[1] Computer Network Information Center, Chinese Academy of Sciences, Beijing, China
[2] University of Chinese Academy of Sciences, Beijing, China
[3] Institute of Pathology, Technical University of Munich, Munich, Germany
[4] School of Biomedical Engineering, Division of Life Sciences and Medicine, University of Science and Technology of China (USTC), Hefei, China
[5] Suzhou Institute for Advanced Research, USTC; Jiangsu Provincial Key Laboratory of Multimodal Digital Twin Technology, Suzhou, China
jjr@sccas.cn, yanrui@ustc.edu.cn

Abstract. Despite notable progress in whole slide image (WSI) analysis, conventional Transformer-based frameworks remain constrained by high computational complexity and limited effectiveness in advanced predictive tasks that demand the joint modeling of global contextual information and fine-grained local features, such as treatment response prediction. This challenge is further exacerbated by the limited sample sizes typically available in prognostic prediction datasets, which hinder robust model generalization. Recent state-space models, such as Mamba, offer improved computational efficiency; however, current Mamba-based designs remain susceptible to information degradation and inherently unidirectional feature propagation, while largely overlooking the multi-scale spatial organization characteristic of pathological images. In this study, we propose MERIT, a hierarchical Multi-scale Mamba framework with Enhanced Retention of Information for Treatment response prediction. Our core contributions include: (1) a dual-branch topology-preserving scanning method that preserves information fidelity and spatial continuity across multiple magnifications; (2) a multi-scale feature fusion module integrating features from multiple multi-scale foundation models via serial Mamba with progressive information inheritance; (3) a retentive information transfer mechanism that mitigates Mamba's long-range forgetting via message token preservation and residual propagation. Evaluated on the Yale breast cancer cohort and the TCGA lung and colorectal datasets, MERIT achieves state-of-the-art AUROCs of 0.880, 0.776, and 0.847, respectively. By consistently outperforming conventional MIL and recent state-space models across all three cohorts, MERIT demonstrates generalizability and advances long-range

[†] Equal Contribution, [⋆] Corresponding Authors.

sequence modeling in precision oncology. Our codes are available at: https://github.com/hutaiyuan/MERIT.

Keywords: Whole slide images · Treatment response prediction.

1 Introduction

Pathological diagnosis is widely regarded as the gold standard for cancer diagnosis, with pathological images containing rich phenotypic information that profoundly influences patient treatment decisions and prognosis evaluation [1]. Advances in digital scanning technologies have enabled the rapid digitization of traditional glass slides into whole slide images (WSIs), thereby significantly enhancing the accessibility and convenience of pathological image analysis [2,3]. WSIs exhibit a multi-resolution hierarchical structure and enormous size, with image dimensions typically reaching the gigapixel level [4]. This massive scale and complexity present significant computational challenges.

Multiple instance learning (MIL) [5] has become one of the widely adopted approaches for analyzing WSIs. It divides WSIs into smaller patches, called instances, which are encoded into low-dimensional features using pretrained models. These features are then aggregated to form a bag-level representation for downstream analysis [6]. MIL treats WSI analysis as a long-sequence modeling task, aiming to capture inter-instance relationships as well as global contextual information to support more effective downstream analysis. Traditional MIL approaches, such as attention-based methods [7,8], often overlook contextual dependencies among instances, limiting the quality of the holistic WSI representation. In response, some studies [9,10] have adopted Transformer [11] to model inter-instance interactions and capture long-range dependencies. However, these methods are often limited by quadratic computational complexity and reduced fitting ability caused by long sequence lengths and limited sample sizes.

Recently, state-space models represented by Mamba have demonstrated efficiency and strong potential in addressing long-range sequence modeling challenges [12,13]. Mamba's low computational complexity and parameter efficiency make it well suited for WSI analysis, where long sequences and limited samples often result in poor fitting [14]. However, Mamba was developed for language tasks and employs a unidirectional causal architecture that propagates information only from past to present tokens. This sequential, one-dimensional design is incompatible with the two-dimensional spatial structure of images, limiting its applicability in image analysis. Recent works such as MambaMIL [15] and 2DMamba [16] have incorporated Mamba into WSI analysis. These methods primarily address input processing without alleviating Mamba's inherent architectural limitations that impact performance. Specifically, its limited parameter capacity and unidirectional causal mechanism lead to a forgetting effect, undermining the modeling of long-range dependencies. Additionally, the lack of hierarchical representation impedes the capture of multi-scale structures essential in WSI analysis. These core limitations continue to constrain the effectiveness of Mamba-based approaches in this domain.

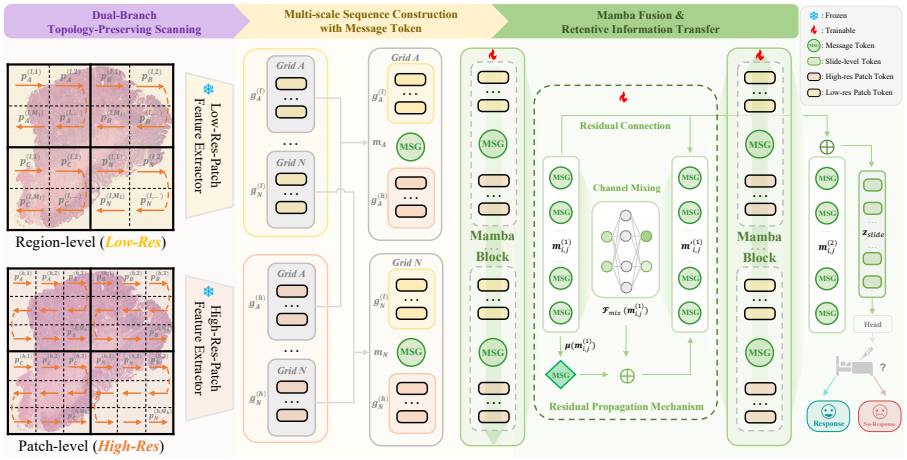

Fig. 1. Overview of MERIT.

We propose **MERIT**, a hierarchical **M**ulti-scale Mamba framework with **E**nhanced **R**etention of **I**nformation for **T**reatment response prediction, facilitating superior information flow and robust long-range dependency modeling. Our main contributions are as follows:

1. To address fragmentation from flattening 2D WSIs into 1D sequences, we propose a dual-branch topology-preserving scanning strategy that preserves multi-scale fidelity and contextual continuity for improved information flow.
2. To capture multi-scale spatial patterns in WSIs, we propose a hierarchical fusion module that integrates features from multiple multi-scale foundation models via serial Mamba with progressive information inheritance, enabling cross-resolution representation learning.
3. To mitigate long-range forgetting caused by Mamba's unique architecture, we propose a novel retentive information transfer mechanism with message token preservation and residual propagation to enhance information retention and long-range context modeling.

In clinical practice, predicting treatment response from pathology images is critical yet remains challenging due to limited sample sizes [17]. We evaluate MERIT across the Yale breast cancer cohort and the TCGA lung and colorectal datasets. The model achieves state-of-the-art AUROCs of 0.880, 0.776, and 0.847 across these three cohorts respectively, surpassing current approaches and demonstrating potential for aiding clinical decisions in precision oncology.

2 Methods

We propose MERIT, a hierarchical multi-scale framework designed to capture long-range dependencies in WSIs while mitigating the forgetting problem inher-

ent in state-space models. As illustrated in Fig. 1, MERIT comprises three integral components: dual-branch topology-preserving scanning across scales, multi-scale sequence construction with message tokens, and serial Mamba featuring retentive information transfer.

2.1 Dual-Branch Topology-Preserving Scanning

Given a WSI $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, we aim to construct a sequence that preserves spatial locality across scales. We partition $\mathbf{I}$ into a grid set $\mathcal{G} = \{G_{i,j}\}_{i,j}^{H_g,W_g}$. To capture multi-granular features, we extract fine-grained high-resolution patches $\mathcal{P}_{i,j}^{(h)}$ and coarse-grained low-resolution patches $\mathcal{P}_{i,j}^{(l)}$ from each grid $G_{i,j}$.

Unlike raster scanning that separates proximal patches and disrupts continuity, we implement *Hierarchical Serpentine Scanning* to preserve local topology. To characterize the serpentine trajectory, we define $\mathcal{J}_k^L$ as an alternating index sequence following $(1,\dots,L)$ for odd k and $(L,\dots,1)$ for even k. For each resolution $r \in \{h,l\}$, we first construct the grid-level embedding $\mathbf{g}_{u,v}^{(r)}$ by concatenating encoded patch features along a local serpentine path within each grid $G_{u,v}$. Subsequently, these grid embeddings are aggregated into a slide-level sequence $\mathbf{S}^{(r)}$ via a global serpentine trajectory:

$$\mathbf{g}_{u,v}^{(r)} = \bigoplus_{i=1}^{h_r} \bigoplus_{j\in\mathcal{J}_i^{w_r}} f_{\text{enc}}^{(r)}(p_{u,v}^{(r,i,j)}), \quad \mathbf{S}^{(r)} = \bigoplus_{u=1}^{H_g} \bigoplus_{v\in\mathcal{J}_u^{W_g}} \mathbf{g}_{u,v}^{(r)}, \tag{1}$$

where $\bigoplus$ denotes sequence concatenation, $f_{\text{enc}}^{(r)}$ is the resolution-specific patch encoder, and $p_{u,v}^{(r,i,j)}$ represents the patch at local coordinate (i,j) within grid (u,v) for resolution r, H_g and W_g denote the total vertical and horizontal grid counts respectively, whereas h_r and w_r represent the patch dimensions within a single grid at resolution r. This strategy ensures that spatially adjacent regions remain contiguous in the 1D sequence $\mathbf{S}^{(r)}$, thereby minimizing the structural distortion caused by flattening.

2.2 Serial Mamba with Enhanced Information Retention

To effectively integrate multi-scale features and mitigate the long-range forgetting inherent in Mamba, we introduce a *Retentive Information Transfer* mechanism inspired by memory consolidation theories [18].

Message Token and Sequence Construction. We introduce a learnable *message token* $\mathbf{m}_{u,v} \in \mathbb{R}^d$ for each grid, acting as a short-term memory unit, storing and integrating local information within the grid, which helps retain critical features across different resolutions. For grid $G_{u,v}$, the composite sequence $\mathbf{s}_{u,v}$ is constructed by positioning a learnable message token at the designated

position $p_{\text{msg}}^{(u,v)}$ between the low-resolution and high-resolution embeddings:

$$\mathbf{s}_{u,v} = \left[\mathbf{g}_{u,v}^{(l)}; \mathbf{m}_{u,v}; \mathbf{g}_{u,v}^{(h)}\right], \quad \mathbf{S} = \bigoplus_{u=1}^{H_g} \bigoplus_{v \in \mathcal{J}_u^{W_g}} \mathbf{s}_{u,v}, \tag{2}$$

Retentive Information Transfer. The unified sequence $\mathbf{S}$ is processed by a serial Mamba module. To enhance feature retention, we propose a *Residual Propagation Mechanism*. Let $\text{SSM}(\cdot)$ denote the Selective State Space Model operation [13]. The process is formulated as a dual-stage update:

Stage 1: Forward Consolidation. We first process the sequence to extract multi-scale contextual features into the message tokens. The token is then updated via a residual mixing function to reinforce memory retention:

$$\begin{aligned} \mathbf{S}' = \text{SSM}_{\text{fwd}}(\mathbf{S}), \quad \mathbf{m}_{i,j}^{(1)} = \mathbf{S}'[p_{\text{msg}}], \\ \mathbf{m}'^{(1)}_{i,j} = \mathbf{m}_{i,j}^{(1)} + \mathcal{F}_{\text{mix}}(\mathbf{m}_{i,j}^{(1)}) + \mu(\mathbf{m}_{i,j}^{(1)}), \end{aligned} \tag{3}$$

where p_{msg} indexes the message token position. The term $\mu(\cdot)$ represents the global mean pooling to bridge the token learning disparity across regions, and $\mathcal{F}_{\text{mix}}(\cdot)$ is a multi-layer perceptron serving as a channel-mixing operation to enhance context modeling. This residual update mimics synaptic reinforcement, facilitating information exchange between different storage units, and enhancing both the retention and interaction of critical features.

Stage 2: Backward Refinement. The updated message tokens, now enriched with contextual and multi-scale information, are re-inserted into the sequence at the same position $p_{\text{msg}}(i,j)$. We then reverse the sequence ($\text{Flip}(\cdot)$) and apply a second Mamba block to capture bidirectional dependencies:

$$\mathbf{S}'' = \text{SSM}_{\text{bwd}}(\text{Flip}(\mathbf{S}'|_{p_{\text{msg}} \leftarrow \mathbf{m}'^{(1)}})), \quad \mathbf{m}_{i,j}^{(2)} = \mathbf{S}''[p_{\text{msg}}]. \tag{4}$$

The final grid representation is the fusion of the consolidated forward and backward memories: $\mathbf{f}_{i,j} = \mathbf{m}'^{(1)}_{i,j} + \mathbf{m}_{i,j}^{(2)}$.

The slide-level representation $\mathbf{z}$ and probability $\hat{y}$ are derived via attention aggregation and linear projection:

$$\mathbf{z} = \sum_{n=1}^{N} \frac{\exp(\mathbf{w}^\top \tanh(\mathbf{V}\mathbf{f}_n^\top))}{\sum_{k=1}^{N} \exp(\mathbf{w}^\top \tanh(\mathbf{V}\mathbf{f}_k^\top))} \mathbf{f}_n, \quad \hat{y} = \sigma(\mathbf{W}_c\mathbf{z} + \mathbf{b}), \tag{5}$$

where $\mathbf{V}$, $\mathbf{w}$, $\mathbf{W}_c$, and $\mathbf{b}$ denote learnable parameters.

3 Experiments

3.1 Dataset

To comprehensively evaluate the proposed MERIT framework, we utilized three public WSI datasets for treatment response prediction.

Table 1. Comparison of MERIT and MIL Methods

Method	Yale-BRCA			TCGA-LUNG			TCGA-CRC		
	Acc.	**AUC**	**F1**	**Acc.**	**AUC**	**F1**	**Acc.**	**AUC**	**F1**
ABMIL	$0.694_{.069}$	$0.843_{.059}$	$0.553_{.140}$	$0.628_{.043}$	$0.686_{.071}$	$0.688_{.052}$	$0.600_{.189}$	$0.523_{.294}$	$0.624_{.239}$
Transformer	$0.718_{.094}$	$0.826_{.068}$	$0.665_{.132}$	$0.658_{.048}$	$0.707_{.060}$	$0.720_{.041}$	$0.629_{.270}$	$0.607_{.327}$	$0.591_{.336}$
TransMIL	$0.729_{.029}$	$\underline{0.865_{.025}}$	$0.601_{.082}$	$0.643_{.073}$	$0.700_{.070}$	$0.698_{.074}$	$\underline{0.686_{.057}}$	$\underline{0.826_{.100}}$	$\underline{0.690_{.146}}$
CLAM	$0.717_{.078}$	$0.862_{.052}$	$0.602_{.139}$	$0.667_{.108}$	$0.706_{.083}$	$0.731_{.078}$	$0.600_{.139}$	$0.756_{.087}$	$0.680_{.184}$
DSMIL	$0.718_{.044}$	$0.854_{.052}$	$0.608_{.078}$	$0.644_{.072}$	$0.705_{.047}$	$0.701_{.061}$	$0.628_{.145}$	$0.743_{.212}$	$0.618_{.315}$
HIPT	$0.682_{.080}$	$0.831_{.067}$	$0.672_{.056}$	$0.651_{.090}$	$0.717_{.097}$	$0.713_{.068}$	$0.601_{.228}$	$0.663_{.309}$	$0.470_{.400}$
2DMamba	$0.671_{.071}$	$0.754_{.112}$	$0.574_{.085}$	$0.613_{.070}$	$0.674_{.146}$	$0.683_{.070}$	$0.514_{.171}$	$0.560_{.303}$	$0.340_{.336}$
SRMamba	$0.694_{.078}$	$0.745_{.087}$	$0.597_{.134}$	$0.659_{.080}$	$0.682_{.093}$	$0.728_{.078}$	$0.657_{.193}$	$0.796_{.280}$	$0.597_{.330}$
MambaMIL	$\underline{0.752_{.114}}$	$0.726_{.197}$	$\underline{0.678_{.142}}$	$0.597_{.047}$	$0.631_{.103}$	$0.690_{.053}$	$0.543_{.245}$	$0.527_{.284}$	$0.559_{.241}$
S4MIL	$0.694_{.024}$	$0.839_{.017}$	$0.550_{.037}$	$\underline{0.674_{.027}}$	$\underline{0.720_{.044}}$	$\underline{0.761_{.022}}$	$0.657_{.145}$	$0.750_{.161}$	$0.647_{.328}$
MERIT	$\mathbf{0.753}_{.107}$	$\mathbf{0.880}_{.036}$	$\mathbf{0.683}_{.153}$	$\mathbf{0.736}_{.050}$	$\mathbf{0.776}_{.047}$	$\mathbf{0.798}_{.047}$	$\mathbf{0.714}_{.090}$	$\mathbf{0.847}_{.141}$	$\mathbf{0.708}_{.163}$

Yale-BRCA Cohort: We analyzed 85 HER2-positive breast cancer cases from the Yale Pathology database. All patients received neoadjuvant targeted therapy prior to surgery. Based on pathology reports, treatment response was categorized into 36 responders and 49 non-responders [19].

TCGA Cohorts: To assess cross-cancer generalizability, we evaluated MERIT on two additional cohorts from The Cancer Genome Atlas (TCGA): TCGA-CRC, including 72 Oxaliplatin-treated colorectal cancer patients with 44 responders and 28 non-responders based on clinical outcomes, and TCGA-LUNG, including 154 lung cancer patients treated with platinum-based chemotherapy (Carboplatin or Cisplatin), with 103 responders and 51 non-responders [20,21].

3.2 Implementation Details

High-resolution patch images were extracted at 20× magnification (0.5 µm/px), while low-resolution patches were sampled at 10× magnification (1 µm/px). All patches were 256×256 pixels, with a grid size of 2560px. For feature extraction, to align with the respective training data scales, UNI2-h [22] was used for 20× patches, and CTransPath [23] for 10× patches. The model was trained using the AdamW optimizer [24] with a learning rate of 2×10^{-5}. Training occurred over 12 epochs with a batch size of 1, using binary cross-entropy loss. Five-fold cross-validation was employed for all training and validation. Experiments were run on a cluster of eight NVIDIA A800 GPUs.

3.3 Evaluation and Comparison of MERIT

As shown in Table 1, we compare MERIT with current leading methods from traditional, multi-scale, and Mamba-based approaches, including ABMIL [7], Transformer [11], TransMIL [25], CLAM [8], DSMIL [26], HIPT [27], 2DMamba [16],

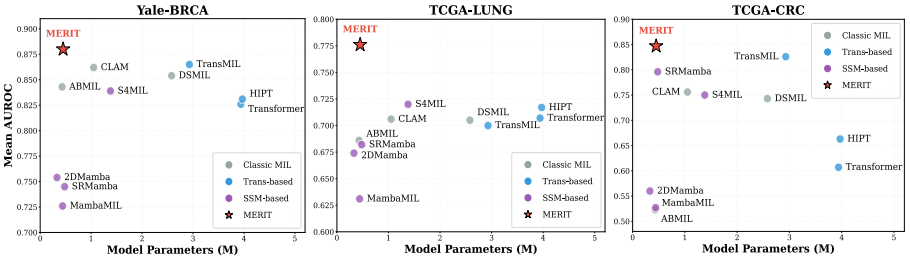

Fig. 2. Efficiency-performance trade-off.

Table 2. Ablation study on key components of MERIT

Variant	Yale-BRCA			TCGA-LUNG			TCGA-CRC		
	Acc.	AUC	F1	Acc.	AUC	F1	Acc.	AUC	F1
w/o C.S.	$0.541_{.146}$	$0.717_{.099}$	$0.474_{.246}$	$0.551_{.143}$	$0.657_{.134}$	$0.493_{.271}$	$0.680_{.193}$	$0.505_{.149}$	$0.689_{.345}$
w/o MSG	$0.729_{.071}$	$0.848_{.063}$	$0.648_{.138}$	$0.675_{.051}$	$0.735_{.066}$	$0.759_{.048}$	$0.560_{.101}$	$0.762_{.167}$	$0.586_{.144}$
w/o R.P.	$0.741_{.060}$	$0.811_{.061}$	$0.679_{.081}$	$0.683_{.093}$	$0.743_{.072}$	$0.770_{.059}$	$0.679_{.132}$	$0.771_{.097}$	$0.708_{.158}$
w/o H.R.	$0.647_{.144}$	$0.603_{.281}$	$0.504_{.267}$	$0.612_{.084}$	$0.643_{.121}$	$0.677_{.130}$	$0.460_{.219}$	$0.676_{.148}$	$0.319_{.436}$
w/o L.R.	$0.753_{.058}$	$0.851_{.043}$	$0.671_{.088}$	$0.652_{.112}$	$0.735_{.061}$	$0.693_{.114}$	$0.600_{.063}$	$0.752_{.097}$	$0.657_{.076}$

SRMamba [15], MambaMIL [15], and S4MIL [28]. To ensure fairness, we followed the original setups and fixed same feature extractors: UNI2-h for 20× patches and CTransPath for 10× patches.

Comparative Performance Analysis: Table 1 shows that MERIT achieves SOTA performance across all cohorts, with AUROCs of 0.880, 0.776, and 0.847 on Yale-BRCA, TCGA-LUNG, and TCGA-CRC, respectively. These results suggest strong predictive capability and cross-cancer generalizability, highlighting its multi-resolution feature preservation and integration for response prediction.

Efficiency Analysis: Figure 2 visualizes the trade-off between performance (Mean AUROC) and complexity (Model Parameters). MERIT consistently occupies the optimal top-left quadrant, achieving SOTA performance with only 0.45M parameters. This high computational efficiency and compact model size underscore its potential for scalable deployment in clinical pathology workflows.

3.4 Ablation Study

An ablation study is provided in Table 2 to evaluate MERIT's core components. Removing Continuous Scanning (C.S.) causes performance degradation, as conventional scanning disrupts local spatial relationships. In contrast, C.S. ensures spatially adjacent grids remain proximal within sequence, ensuring spatial continuity. This significant gap validates that continuity is indispensable for modeling short-range cellular dependencies and multi-scale morphological features. Excluding the Message Token (MSG) or Residual Propagation (R.P.) causes per-

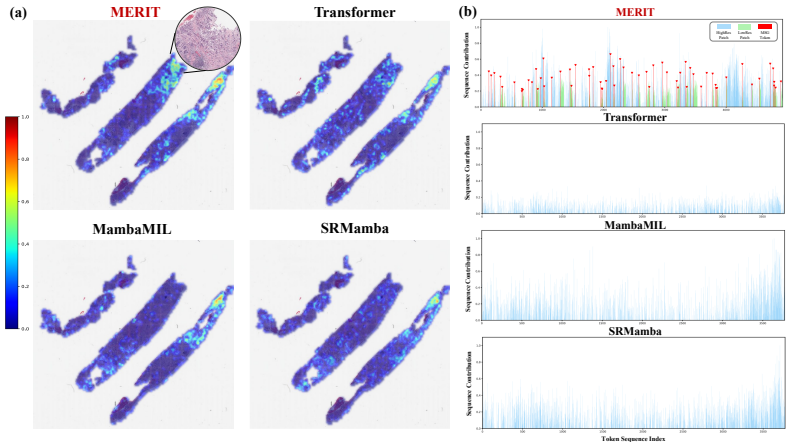

Fig. 3. (a) Grad-CAM heatmaps and (b) Sequence contribution.

formance drops. MSG acts as a crucial short-term memory unit that stores and progressively transfers local contextual information across scales. While R.P. facilitates information exchange and transfer, analogous to biological long-term memory consolidation. The performance drops in both variants empirically confirm that the integration of MSG and R.P. is essential for enhancing long-range dependency modeling and retaining vital clinical features. Performance drops upon excluding either High-Resolution (H.R.) or Low-Resolution (L.R.) features validate our hierarchical design. Specifically, omitting H.R. inputs induces the severe accuracy collapse, identifying fine-grained cellular details as primary predictive drivers; meanwhile, the decline without L.R. inputs underscores the necessity of region-level context for essential spatial grounding.

3.5 Interpretability Analysis

We evaluate MERIT's mechanisms through spatial and sequence-level analysis (Fig. 3), revealing its superior spatial fidelity, robust information retention, and the MSG Token's efficacy as an information hub. MERIT's Grad-CAM[29] highlights clinically discriminative regions, such as tumor-rich regions identified by pathologists. Notably, MERIT identifies effective tissue patterns overlooked by existing methods, confirming that our design effectively preserves fine-grained morphological details essential for prediction. We assess long-range dependencies by plotting Sequence Contribution across the Token Sequence Index. While Transformers exhibit diluted focus and Mamba variants suffer from unidirectional forgetting, MERIT ensures robust global retention by mitigating information decay while selectively prioritizing key pathological regions, effectively balancing global context with local salience. The MSG Token acts as a pivotal hub with peak contribution, bridging cross-resolution patches to effectively capture comprehensive multi-scale patterns.

4 Conclusion

In this work, we introduce MERIT, a multi-scale Mamba framework for WSI that overcomes key limitations of existing Mamba-based models, including information forgetting and unidirectional learning. By integrating a dual-branch scanning strategy to preserve multi-scale contextual continuity, a hierarchical fusion module with serial Mamba for cross-resolution learning, and a novel retentive information transfer mechanism to mitigate long-range forgetting, our method effectively enhances information retention and context dependency modeling. Experiments across three clinical cohorts demonstrate that our approach outperforms state-of-the-art methods, showing potential for clinical treatment response prediction.

Acknowledgments This work was supported by the Noncommunicable Chronic Diseases-National Science and Technology Major Project (No. 2024ZD0533500), National Natural Science Foundation of China (No. 62402473), and Basic Research Program of Jiangsu (No.BK20255001).

Disclosure of Interests The authors have no competing interests to declare that are relevant to the content of this article.

References

1. El Nahhas, O.S., van Treeck, M., Wölflein, G., Unger, M., Ligero, M., Lenz, T., Wagner, S.J., Hewitt, K.J., Khader, F., Foersch, S., et al.: From whole-slide image to biomarker prediction: end-to-end weakly supervised deep learning in computational pathology. Nature protocols **20**(1), 293–316 (2025)
2. Luan, H., Hu, T., Hu, J., Li, R., Ji, D., He, J., Duan, X., Yang, C., Gao, Y., Chen, F., et al.: Multi-class cancer classification of whole slide images through transformer and multiple instance learning. In: International symposium on bioinformatics research and applications. pp. 150–164. Springer (2023)
3. Claudio Quiros, A., Coudray, N., Yeaton, A., Yang, X., Liu, B., Le, H., Chiriboga, L., Karimkhan, A., Narula, N., Moore, D.A., et al.: Mapping the landscape of histomorphological cancer phenotypes using self-supervised learning on unannotated pathology slides. Nature Communications **15**(1), 4596 (2024)
4. Wang, B., Zhang, K., Wang, Y., Gu, Y., Luan, H., Zhou, Y., Hu, T., Wang, R., Yang, Z., Jiang, Z., et al.: Wsisum: Wsi summarization via dual-level semantic reconstruction. Medical Image Analysis p. 103970 (2026)
5. Carbonneau, M.A., Cheplygina, V., Granger, E., Gagnon, G.: Multiple instance learning: A survey of problem characteristics and applications. Pattern recognition **77**, 329–353 (2018)
6. Luan, H., Yang, K., Hu, T., Hu, J., Liu, S., Li, R., He, J., Yan, R., Guo, X., Qian, N., et al.: Review of deep learning-based pathological image classification: From task-specific models to foundation models. Future Generation Computer Systems **164**, 107578 (2025)
7. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018)

8. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature biomedical engineering **5**(6), 555–570 (2021)
9. Luan, H., Hu, T., Hu, J., Liu, W., Yang, K., Pei, Y., Li, R., He, J., Gao, Y., Sun, D., et al.: Breast cancer homologous recombination deficiency prediction from pathological images with a sufficient and representative transformer. NPJ Precision Oncology **9**(1), 160 (2025)
10. Hu, T., Luan, H., Yan, R., Hu, J., Yang, K., Han, X., Liu, W., He, J., Duan, X., Li, R., et al.: Transformer-based multi-scale fusion for robust predicting microsatellite instability from pathological images. In: 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 2046–2053. IEEE (2024)
11. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
12. Gu, A., Goel, K., Ré, C.: Efficiently modeling long sequences with structured state spaces. arXiv preprint arXiv:2111.00396 (2021)
13. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
14. Yang, Z., Shi, X., Ba, W., Song, Z., Luan, H., Hu, T., Lin, S., Wang, J., Zhou, S.K., Yan, R.: Fusion of multi-scale heterogeneous pathology foundation models for whole slide image analysis. arXiv preprint arXiv:2510.27237 (2025)
15. Yang, S., Wang, Y., Chen, H.: Mambamil: Enhancing long sequence modeling with sequence reordering in computational pathology. In: International conference on medical image computing and computer-assisted intervention. pp. 296–306. Springer (2024)
16. Zhang, J., Nguyen, A.T., Han, X., Trinh, V.Q.H., Qin, H., Samaras, D., Hosseini, M.S.: 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 (2025)
17. Wang, X., Zhao, J., Marostica, E., Yuan, W., Jin, J., Zhang, J., Li, R., Tang, H., Wang, K., Li, Y., et al.: A pathology foundation model for cancer diagnosis and prognosis prediction. Nature **634**(8035), 970–978 (2024)
18. Morrison, J.H., Baxter, M.G.: The ageing cortical synapse: hallmarks and implications for cognitive decline. Nature Reviews Neuroscience **13**(4), 240–250 (2012)
19. Farahmand, S., Fernandez, A.I., Ahmed, F.S., Rimm, D.L., Chuang, J.H., Reisenbichler, E., Zarringhalam, K.: Her2 and trastuzumab treatment response h&e slides with tumor roi annotations (version 3) (2022), https://www.cancerimagingarchive.net/, [Data set]
20. Liu, J., Lichtenberg, T., Hoadley, K.A., Poisson, L.M., Lazar, A.J., Cherniack, A.D., Kovatich, A.J., Benz, C.C., Levine, D.A., Lee, A.V., et al.: An integrated tcga pan-cancer clinical data resource to drive high-quality survival outcome analytics. Cell **173**(2), 400–416 (2018)
21. Dalibor Hrg, Balthasar Huber, L.A.H.: Tcga chemotherapy response dataset (v1.0) (2020), https://doi.org/10.5281/zenodo.3719291, [Data set]
22. 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.: Towards a general-purpose foundation model for computational pathology. Nature Medicine (2024)
23. Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., Huang, J., Han, X.: Transformer-based unsupervised contrastive learning for histopathological image classification. Medical image analysis **81**, 102559 (2022)

24. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
25. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. Advances in neural information processing systems **34**, 2136–2147 (2021)
26. Li, B., Li, Y., Eliceiri, K.W.: 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. pp. 14318–14328 (2021)
27. Chen, R.J., Chen, C., Li, Y., Chen, T.Y., Trister, A.D., Krishnan, R.G., Mahmood, F.: Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16144–16155 (2022)
28. Fillioux, L., Boyd, J., Vakalopoulou, M., Cournède, P.H., Christodoulidis, S.: Structured state space models for multiple instance learning in digital pathology. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 594–604. Springer (2023)
29. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)