AdaSurvMamba: Dynamic Fusion and Semantic Scanning for Multimodal Survival Analysis

Jialong Zhong[1], Tingwei Liu[1], Baokun Yue[1], Jingjing Li[2], Yongri Piao[1(✉)], Miao Zhang[1(✉)], Leiye Liu[1], Jiahong Jiang[1], Wei Ji[3], and Huchuan Lu[1]

[1] Dalian University of Technology
[2] University of Alberta
[3] Yale University
yrpiao@dlut.edu.cn, miaozhang@dlut.edu.cn

Abstract. Multimodal survival analysis utilizing whole slide images (WSIs) and genomic profiles is fundamental for cancer prognosis. Recently, state-space models like Mamba have emerged as powerful tools for sequence modeling. However, translating this success to complex multimodal tasks is hindered by two critical limitations. First, conventional fusion strategies assume a static multimodal interaction strength, ignoring the fluctuating diagnostic importance of each modality across different patients and local regions. Second, the standard Mamba architecture processes tokens along predefined physical paths. This rigid scanning disrupts the semantic continuity of spatially scattered medical features and exacerbates long-range decay. To address these challenges, we introduce AdaSurvMamba as a novel adaptive framework for multimodal survival analysis. The framework features a Dual-Scale Importance-Aware Reconstruction (DSIR) module to dynamically modulate cross-modal interaction strength. It evaluates diagnostic importance at both the sequence and token levels to reconstruct the input representations. Furthermore, we propose a Semantic Aggregation Scanning (SAS) module to overcome contextual fragmentation. The SAS module dynamically reorganizes discrete tokens into semantically continuous sequences via a shared prototype pool. It explicitly modulates the state transition step size using global modality context and semantic priors to adaptively control the information absorption rate. Experiments across five TCGA cohorts demonstrate consistent gains over existing methods. Code is available at https://github.com/zjlGO/AdaSurvMamba.

Keywords: Survival outcome prediction · Whole slide image · Mamba.

1 Introduction

Deep learning has greatly advanced WSI analysis in computational pathology [26, 28]. Survival analysis is a fundamental task in cancer prognosis and treatment planning [20]. Accurate prediction requires a comprehensive understanding of the tumor, which further motivates the use of multimodal data in clinical decision-making. Specifically, Whole Slide Images (WSIs) reveal detailed tumor

morphology, while genomic profiles provide critical insights into molecular alterations. These two modalities offer distinct yet highly complementary prognostic clues [19,21]. Consequently, multimodal deep learning approaches that integrate WSIs and genomics have recently achieved remarkable success in survival prediction [3,4,7].

Many early approaches in this field rely on late fusion to integrate these diverse features [2,4]. While effective to some extent, these strategies often fail to capture critical cross-modal synergies. Recent methods have attempted to address this limitation by introducing attention mechanisms or hybrid supervision to better align pathological and genomic features [3,7,8,17]. In parallel, the Mamba architecture has achieved notable success in sequence modeling tasks, such as whole slide image classification [12,24,25,27]. This success has inspired initial attempts to apply Mamba to multimodal survival analysis, leading to frameworks like SurvMamba [5] and ME-Mamba [23]. Unlike prior survival models using fixed cross-modal interactions or predefined scans, DSIR learns patient- and token-level modality relevance, while SAS performs semantic reordering before scanning. This also differs from visual Mamba variants that mainly redesign spatial scan paths for regular images, as our design targets sparse WSI regions and unordered genomic features.

Another challenge is the disruption of semantic continuity during the scanning process. Standard Mamba models process features along predefined scanning paths. However, the specific arrangement of input tokens profoundly impacts the sequence modeling performance [11,16]. In WSIs, biologically related tissue patches often scatter across distant locations. Genomic features also lack a natural sequential order. Processing these inputs in a fixed order forces semantically connected elements to be far apart in the sequence. Since Mamba suffers from long-range decay [9], this forced distance prevents effective interactions and causes severe semantic fragmentation. Consequently, how to gather these discrete features into semantically connected sequences remains a critical problem to address.

To address these challenges, we propose AdaSurvMamba, a novel adaptive framework for multimodal survival analysis. It dynamically controls feature interactions and maintains semantic continuity. The framework contains a Dual-Scale Importance-Aware Reconstruction (DSIR) module and a Semantic Aggregation Scanning (SAS) module. The DSIR module overcomes the rigidity of existing fusion methods. It evaluates diagnostic importance at both the sequence and token levels to rebuild the input representations. This reconstruction directly regulates the multimodal interaction strength. It allows the model to focus on the most critical prognostic signals. Building upon this, the SAS module overcomes the issue of semantic fragmentation. It further reorganizes the rebuilt features into semantically continuous scanning sequences. This keeps related tokens connected during the state transition process and avoids the drawbacks of predefined scanning paths. Together, the DSIR and SAS modules enable AdaSurvMamba to flexibly integrate features and capture robust context. Extensive experiments on multiple public datasets validate the effectiveness of our framework.

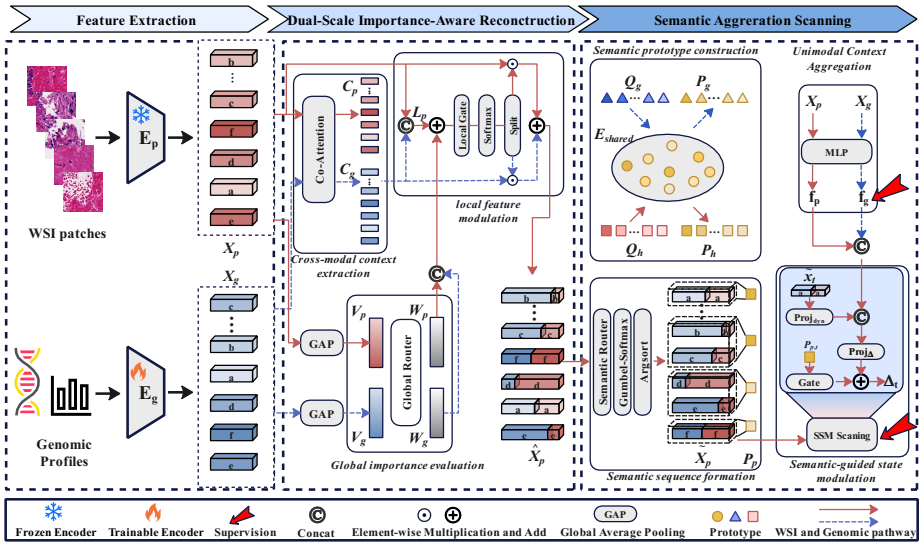

Fig. 1. Overview of the proposed AdaSurvMamba framework. For visual clarity, the detailed computational flow within the DSIR and SAS modules is primarily illustrated using the pathology (WSI) branch, while the genomic branch follows a symmetric processing paradigm.

2 Method

Figure 1 illustrates the overall architecture of AdaSurvMamba. The framework processes WSI patches and genomic profiles. A frozen foundation model encodes WSI patches, while a learnable SNN maps six functional groups derived from RNA-seq, CNV, and mutation data into six tokens $\mathbf{X}_g \in \mathbb{R}^{6\times D}$. The Dual-Scale Importance-Aware Reconstruction (DSIR) module then computes global and local importance weights to dynamically adjust these multimodal features. Next, the Semantic Aggregation Scanning (SAS) module reorganizes the adjusted tokens into semantically coherent sequences for state space modeling. Finally, the network aggregates the enhanced representations to predict patient survival.

2.1 Preliminaries

Mamba formulates sequence modeling through structured state space models (SSMs). The continuous dynamics map an input $x(t)$ to output $y(t)$ via a hidden state $h(t)$:

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

where $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{D}$ are learnable matrices. Discretization converts this process into a recurrent form:

$$h_t = \bar{\mathbf{A}}h_{t-1} + \bar{\mathbf{B}}x_t. \tag{2}$$

The key innovation lies in its selectivity mechanism. The parameters $\mathbf{B}$, $\mathbf{C}$, and the timescale Δ become input-dependent functions. This allows the model to selectively retain or forget information during state transitions. However, this recurrent process strictly follows the predefined physical order of input tokens. Such rigidity prevents effective modeling of long-range dependencies in multimodal medical data.

2.2 Dual-Scale Importance-Aware Reconstruction

Existing fusion methods apply static interaction weights across all samples. This overlooks the fact that modality importance varies both across patients and within local feature pairs. The DSIR module addresses this by dynamically reconstructing features at two complementary scales.

Cross-modal context extraction. We establish early interactions between pathological tokens $\mathbf{X}_p \in \mathbb{R}^{N\times D}$ and genomic tokens $\mathbf{X}_g \in \mathbb{R}^{M\times D}$. To avoid the heavy computational burden of standard cross-attention, we employ an efficient co-attention mechanism. It calculates a joint affinity matrix from the linearly projected modalities to derive attention vectors, which then explicitly modulate the original tokens. This produces context-aware features $\mathbf{C}_p \in \mathbb{R}^{N\times D}$ and $\mathbf{C}_g \in \mathbb{R}^{M\times D}$, effectively bridging the initial semantic gap with minimal overhead.

Global importance evaluation. The dominant prognostic modality often differs across patients. We quantify this macroscopic preference through a global confidence router. Global average pooling condenses $\mathbf{X}_p$ and $\mathbf{X}_g$ into vectors $\mathbf{v}_p, \mathbf{v}_g \in \mathbb{R}^D$. A multi-layer perceptron processes their concatenation to produce sequence-level importance weights $\mathbf{w}_p, \mathbf{w}_g \in \mathbb{R}^D$. These weights reflect the overall contribution of each modality for the current patient.

Local feature modulation. Even after global adjustment, the importance of individual token pairs still varies. We model this microscopic variation through a local fusion network. Taking the pathological branch as an example, the network processes the concatenated tensor $[\mathbf{X}_p, \mathbf{C}_g]$ to produce initial gate logits $\mathbf{L}_p \in \mathbb{R}^{N\times 2\times D}$. We then inject the global importance weights to obtain the final gating signals:

$$\tilde{\mathbf{L}}_p[:,0,:] = \mathbf{L}_p[:,0,:] + \mathbf{w}_p, \quad \tilde{\mathbf{L}}_p[:,1,:] = \mathbf{L}_p[:,1,:] + \mathbf{w}_g. \tag{3}$$

A softmax operation applied along the modality dimension yields the dynamic gates $\mathbf{G}_{raw}$ and $\mathbf{G}_{ctx}$. The reconstructed pathological representation becomes:

$$\hat{\mathbf{X}}_p = \mathbf{X}_p \odot \mathbf{G}_{raw} + \mathbf{C}_g \odot \mathbf{G}_{ctx}. \tag{4}$$

The genomic branch undergoes a symmetric process to produce $\hat{\mathbf{X}}_g \in \mathbb{R}^{M\times D}$. This dual-scale strategy amplifies critical signals before they enter the SSM.

2.3 Semantic Aggregation Scanning

Standard Mamba processes tokens along fixed physical paths. This disrupts the semantic continuity of medical features that are spatially scattered in the original data. The SAS module solves this by dynamically reorganizing tokens and modulating state transitions based on semantic content.

Semantic prototype construction. Pathology and genomics have distinct information structures. We build a shared prototype pool $\mathbf{E}_{shared} \in \mathbb{R}^{r \times d_p}$ to map both modalities into a unified semantic space. Here r is the rank for semantic decoupling and d_p is the prototype dimension. For modality $m \in \{p, g\}$, we learn a coefficient matrix $\mathbf{Q}_m \in \mathbb{R}^{T_m \times r}$ to extract modality-specific prototypes:

$$\mathbf{P}_m = \mathbf{Q}_m \mathbf{E}_{shared} \in \mathbb{R}^{T_m \times d_p}, \tag{5}$$

where T_m is the number of prototypes for that modality. This design captures modal specifics while enforcing cross-modal alignment through the shared pool. Since genomics is compressed into six pathway-level tokens, fewer prototypes summarize major molecular programs without over-fragmentation, whereas numerous heterogeneous WSI patches require finer grouping. This is an ablation-supported design choice rather than an assumption of limited genomic heterogeneity.

Semantic sequence formation. A linear router and Gumbel-Softmax generate discrete routing weights $\mathbf{R}_m \in \mathbb{R}^{L_m \times T_m}$ for each token in $\hat{\mathbf{X}}_m$. These weights serve a dual architectural purpose. Spatially, they dictate an argsort operation to deterministically cluster identically assigned tokens, forming the continuous sequence $\tilde{\mathbf{X}}_m$. Semantically, they are multiplied with the prototype pool to extract a differentiable context bias. By explicitly injecting this bias into the continuous SSM step size (Δ_t), the model establishes a valid gradient pathway back to the router, thereby optimizing the discrete semantic reorganization end-to-end.

Unimodal Context Aggregation. Before entering the SAS module, each modality branch independently aggregates its reconstructed sequence via attention pooling. This produces global feature vectors $\mathbf{f}_p \in \mathbb{R}^D$ and $\mathbf{f}_g \in \mathbb{R}^D$, which capture the overall diagnostic signal of pathology and genomics for the current patient. These vectors are also passed through a classifier to generate auxiliary hazard predictions $\hat{\mathbf{h}}_p$ and $\hat{\mathbf{h}}_g$, providing additional supervision. The global features serve as compact representations of modality-level importance.

Semantic-guided state modulation. In state space models, the step size Δ governs how much new information flows into the hidden state. We explicitly guide this flow using both the semantic prototypes and the global modality context. First, we compress each token feature $\hat{\mathbf{x}}_t$ (the t-th row of $\tilde{\mathbf{X}}_m$) into a low-dimensional dynamic vector $\mathbf{x}_t^{\text{dyn}}$ via a linear projection Proj_{dyn}:

$$\mathbf{x}_t^{\text{dyn}} = \text{Proj}_{\text{dyn}}(\hat{\mathbf{x}}_t) \in \mathbb{R}^{d_{\text{dyn}}}. \tag{6}$$

We then concatenate $\mathbf{x}_t^{\text{dyn}}$ with the global features $\mathbf{f} = [\mathbf{f}_p; \mathbf{f}_g] \in \mathbb{R}^{2D}$ and feed the result into a step size generator Proj_{Δ}, a linear layer that outputs the base step size:

$$\Delta_{\text{base},t} = \text{Proj}_{\Delta}\big([\mathbf{x}_t^{\text{dyn}}; \mathbf{f}]\big) \in \mathbb{R}^{d_\Delta}. \tag{7}$$

This injects patient-level modality dominance into the token-wise state transition. Meanwhile, we compute a prototype bias by aggregating the modality-specific prototypes according to the routing weights:

$$\mathbf{B}_{\text{proto},t} = \text{Gate}\big(\mathbf{R}_{m,t}\mathbf{P}_m\big) \in \mathbb{R}^{d_\Delta}, \tag{8}$$

where $\mathbf{R}_{m,t} \in \mathbb{R}^{T_m}$ is the routing vector for token t, and Gate is a linear layer followed by tanh activation. The final semantic-guided step size is the sum of both contributions:

$$\Delta_t = \Delta_{\text{base},t} + \mathbf{B}_{\text{proto},t}. \tag{9}$$

We apply the selective scan using Δ_t on the reordered sequence $\tilde{\mathbf{X}}_m$. This ensures that tokens belonging to the same semantic group receive coherent state updates, and that the update intensity is further modulated by the overall importance of each modality. Following the scan, the enhanced token representations are aggregated into a unified patient profile and fed into a prediction head to estimate the discrete-time survival hazard probabilities.

Optimization objective. Following [3], we apply discrete-time negative log-likelihood to the fused and two unimodal predictions. The total objective is $\mathcal{L} = \mathcal{L}_{\text{main}} + \lambda(\mathcal{L}^p_{\text{aux}} + \mathcal{L}^g_{\text{aux}})$, where the auxiliary terms preserve modality-specific prognostic information.

3 Experiments

3.1 Experiment Protocol

Datasets. We evaluated AdaSurvMamba on five TCGA [14] benchmark cohorts: BLCA ($n = 373$), BRCA ($n = 956$), GBMLGG ($n = 569$), LUAD ($n = 453$), and UCEC ($n = 480$). Each patient comprises diagnostic WSIs and paired genomic profiles (RNA-Seq, copy number variation, and somatic mutations). Following established protocols, we grouped the genomic features into six functional categories.

Implementation details. WSIs were cropped into 512×512 patches at $20\times$ magnification. Pre-trained foundation models (ResNet-50 [10] and UNI [1]) extracted the instance-level morphological features. For the proposed SAS module, the semantic decoupling rank r and the prototype dimension d_p were set to 32 and 64, respectively. Based on our ablation studies, the optimal number of prototypes for genomics and pathology were configured as $T_g = 4$ and $T_p = 16$. We implemented the framework in PyTorch on a single RTX 4090 GPU. The network was optimized using Adam with a learning rate of 2×10^{-4} and weight decay of 1×10^{-5}. The auxiliary loss weight λ was empirically set to 0.5 to balance fused survival prediction and modality-specific supervision. We report the mean Concordance Index (C-Index) and standard deviation from 5-fold cross-validation.

3.2 Comparison with State-of-the-Arts

Table 1 presents the performance comparison of AdaSurvMamba against existing methods across five TCGA cohorts. We include both unimodal baselines (using only genomics or only WSIs) and state-of-the-art multimodal fusion approaches. For genomic-only methods, we compare with MLP and SNN [15]. For WSI-only methods, we consider ABMIL [13] and Transmil [18]. For multimodal

Table 1. Results of different methods. The best results are shown in bold, and the second-best results are underlined.

Method	Modality	BLCA	BRCA	LUAD	GBMLGG	UCEC
Genomic Profiles						
MLP	Genomics	0.611±0.030	0.619±0.053	0.619±0.044	0.779±0.035	0.635±0.069
SNN [15]	Genomics	0.633±0.074	0.634±0.051	0.622±0.041	0.816±0.028	0.620±0.039
Feature extractor: ResNet50 [10]						
ABMIL [13]	WSI	0.589±0.058	0.577±0.084	0.567±0.042	0.771±0.033	0.668±0.019
Transmil [18]	WSI	0.579±0.046	0.640±0.035	0.585±0.039	0.705±0.094	0.597±0.086
MCAT [3]	Multimodal	0.677±0.036	0.657±0.048	0.662±0.035	0.847±0.042	0.675±0.045
MOTCAT [22]	Multimodal	0.677±0.024	0.677±0.020	0.678±0.033	0.841±0.015	0.672±0.042
MMP [19]	Multimodal	0.626±0.042	0.608±0.035	0.650±0.049	0.816±0.021	0.648±0.069
CCL [29]	Multimodal	0.652±0.038	0.596±0.054	0.640±0.057	0.845±0.014	0.686±0.046
GHANT [21]	Multimodal	0.605±0.016	0.674±0.071	0.595±0.059	0.800±0.024	0.652±0.039
LD-VAE [30]	Multimodal	0.614±0.026	0.632±0.065	0.638±0.035	0.841±0.026	0.694±0.046
MGCM [6]	Multimodal	0.675±0.024	0.681±0.057	0.669±0.059	0.845±0.024	0.688±0.039
Ours	Multimodal	**0.707±0.029**	**0.698±0.049**	**0.691±0.040**	**0.859±0.023**	**0.707±0.046**
Feature extractor: UNI [1]						
ABMIL [13]	WSI	0.611±0.034	0.630±0.043	0.640±0.016	0.803±0.047	0.644±0.066
Transmil [18]	WSI	0.590±0.073	0.642±0.037	0.561±0.066	0.735±0.048	0.634±0.064
MCAT [3]	Multimodal	0.682±0.052	0.667±0.023	0.671±0.036	0.842±0.016	0.688±0.041
MOTCAT [22]	Multimodal	0.678±0.026	0.690±0.019	0.673±0.028	0.840±0.023	0.690±0.050
MMP [19]	Multimodal	0.611±0.029	0.592±0.016	0.648±0.035	0.815±0.023	0.654±0.056
CCL [29]	Multimodal	0.657±0.034	0.620±0.051	0.649±0.056	0.852±0.032	0.682±0.050
GHANT [21]	Multimodal	0.622±0.042	0.667±0.029	0.608±0.059	0.808±0.016	0.705±0.036
LD-VAE [30]	Multimodal	0.650±0.015	0.673±0.026	0.643±0.057	0.833±0.025	0.696±0.053
MGCM [6]	Multimodal	0.695±0.024	0.689±0.027	0.679±0.033	0.850±0.021	0.702±0.039
Ours	Multimodal	**0.723±0.026**	**0.701±0.039**	**0.696±0.032**	**0.866±0.017**	**0.718±0.038**

methods, we evaluate against advanced frameworks including MCAT [3], MOTCAT [22], MMP [19], CCL [29], GHANT [21], LD-VAE [30], and MGCM [6]. SurvMamba [5] and ME-Mamba [23] are not included because their official implementations were not publicly available at submission time, preventing reproduction under our unified protocol. As shown in Table 1, AdaSurvMamba achieves consistent improvements over these competitors across all datasets. Specifically, under the UNI setting, our framework achieves an average C-Index improvement of 2.32% over the best competing method across all five cohorts. Similar gains are observed with ResNet50 features, where the average improvement reaches 2.43%. These consistent performance gains highlight the effectiveness of our adaptive multimodal interaction strategy for accurate survival prediction.

3.3 Ablation Study

We conduct extensive ablation studies on the BLCA and BRCA cohorts using ResNet50 features to systematically validate our proposed designs. The baseline model is a standard Mamba architecture utilizing straightforward concatenation fusion and default 1D physical scanning.

Effectiveness of Core Modules. Table 2 isolates the macroscopic contributions of our architectural components. The baseline yields suboptimal performance due to its static fusion and rigid scanning paths. Integrating the DSIR module notably improves the C-Index, demonstrating the necessity of dynamically modulating multimodal interactions. Furthermore, employing the SAS module effectively mitigates long-range decay by reorganizing scattered features into semantically continuous sequences, achieving the best overall performance when combined.

Internal Mechanisms of DSIR. We investigate the fine-grained design of the DSIR module in Table 3. We compare our global-guided dynamic gating against naive modulation strategies. Our gating mechanism significantly outperforms simple scalar shifting or additive attention. This indicates that a simple spatial shift is insufficient to model microscopic token-pair relevance, whereas our softmax-based gating decisively amplifies critical local prognostic signals under macroscopic guidance.

Internal Mechanisms of SAS. We deeply analyze the SAS module from two perspectives: scanning strategy and semantic capacity. First, Table 4 compares our semantic scanning against alternative scanning paths, including conventional 1-way (sequential) and 2-way (bidirectional) physical scans, as well as a random shuffle scan. While 2-way scanning marginally improves upon the 1-way baseline by capturing bidirectional physical contexts, it still suffers from spatial rigidity. Conversely, the random shuffle strategy severely disrupts inherent local continuity, leading to performance degradation. Our semantic scanning, however, leverages learned prototypes to meaningfully reorganize tokens. It significantly outperforms all physical and random scanning paths by constructing semantically coherent sequences, while requiring only a single efficient forward pass. Second, Table 5 analyzes the impact of modality-specific prototype sizes (T_g, T_p). The best setting $(T_g = 4, T_p = 16)$ reflects the different token scales rather than limited genomic heterogeneity.

Table 2. Ablation of core modules.

DSIR	SAS	BLCA	BRCA
-	-	0.642±0.035	0.645±0.041
✓	-	0.675±0.032	0.668±0.045
-	✓	0.669±0.028	0.661±0.038
✓	✓	**0.707±0.029**	**0.698±0.049**

Table 3. DSIR modulation strategies.

Strategy	BLCA	BRCA
Concat (No DSIR)	0.669±0.028	0.661±0.038
Additive Attn	0.681±0.034	0.674±0.041
Simple Scalar	0.693±0.031	0.685±0.045
DSIR Gating	**0.707±0.029**	**0.698±0.049**

Table 4. SAS scanning paths.

Scanning Path	BLCA	BRCA
1-way Physical	0.675±0.032	0.668±0.045
2-way Physical	0.684±0.030	0.679±0.042
Random Shuffle	0.661±0.036	0.655±0.050
Semantic Scan (Ours)	**0.707±0.029**	**0.698±0.049**

Table 5. SAS prototype sizes (T_g, T_p).

Sizes (T_g, T_p)	BLCA	BRCA
(4, 4)	0.688±0.031	0.681±0.044
(4, 8)	0.697±0.030	0.692±0.045
(4, 16)	**0.707±0.029**	**0.698±0.049**
(8, 8)	0.692±0.028	0.686±0.045

4 Conclusion

We presented AdaSurvMamba, an adaptive state-space framework for multimodal cancer survival prediction. By introducing DSIR for dynamic interaction modulation and SAS for semantic scanning with adaptive step-size control, our approach overcomes the limitations of static fusion and rigid physical scanning. Results on five TCGA cohorts show consistent gains. External validation remains future work because cohorts with matched WSIs, genomic profiles, and survival outcomes are still scarce.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China under Grant 62376050, Grant 62372080, in part by the Natural Science Foundation of Liaoning Province under Grant 2024-MSBA-24, and in part by the Dalian Science and Technology Innovation Foundation (2023JJ11CG001).

Disclosure of Interests. We declared no competing interests.

References

1. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., et al.: Towards a general-purpose foundation model for computational pathology. Nature medicine **30**(3), 850–862 (2024)
2. Chen, R.J., Lu, M.Y., Wang, J., Williamson, D.F., Rodig, S.J., Lindeman, N.I., Mahmood, F.: Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. IEEE transactions on medical imaging **41**(4), 757–770 (2020)
3. Chen, R.J., Lu, M.Y., Weng, W.H., Chen, T.Y., Williamson, D.F., Manz, T., Shady, M., Mahmood, F.: Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4025 (2021)
4. Chen, R.J., Lu, M.Y., Williamson, D.F., Chen, T.Y., Lipkova, J., Noor, Z., Shaban, M., Shady, M., Williams, M., Joo, B., et al.: Pan-cancer integrative histology-genomic analysis via multimodal deep learning. Cancer cell **40**(8), 865–878 (2022)
5. Chen, Y., Xie, J., Lin, Y., Song, Y., Yang, W., Yu, R.: Survmamba: state space model with multi-grained multi-modal interaction for survival prediction. arXiv preprint arXiv:2404.08027 (2024)
6. Cui, J., Li, Y., Shen, D., Wang, Y.: Mgcm: Multi-modal graph convolutional mamba for cancer survival prediction. Pattern Recognition **169**, 111991 (2026)
7. Ding, S., Li, J., Wang, J., Ying, S., Shi, J.: Multimodal co-attention fusion network with online data augmentation for cancer subtype classification. IEEE Transactions on Medical Imaging **43**(11), 3977–3989 (2024)
8. Fu, B., He, J., Zhang, X., Peng, Y., Zhang, Z., Tang, Q., Liu, X., Cao, Y., Wang, R.: Hsfsurv: A hybrid supervision framework at individual and feature levels for multimodal cancer survival analysis. Medical Image Analysis p. 103810 (2025)
9. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. In: First conference on language modeling (2024)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)

11. Huang, T., Pei, X., You, S., Wang, F., Qian, C., Xu, C.: Localmamba: Visual state space model with windowed selective scan. In: European conference on computer vision. pp. 12–22. Springer (2024)
12. Huang, Y., Zhao, W., Fu, Y., Zhu, L., Yu, L.: Unleash the power of state space model for whole slide image with local aware scanning and importance resampling. IEEE Transactions on Medical Imaging **44**(2), 1032–1042 (2024)
13. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018)
14. Kandoth, C., McLellan, M.D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J.F., Wyczalkowski, M.A., et al.: Mutational landscape and significance across 12 major cancer types. Nature **502**(7471), 333–339 (2013)
15. Klambauer, G., Unterthiner, T., Mayr, A., Hochreiter, S.: Self-normalizing neural networks. Advances in neural information processing systems **30** (2017)
16. Liu, L., Zhang, M., Yin, J., Liu, T., Ji, W., Piao, Y., Lu, H.: Defmamba: Deformable visual state space model. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 8838–8847 (2025)
17. Liu, M., Cai, C., Li, J., Xu, P., Li, J., Ma, J., Xu, J.: Murrenet: Modeling holistic multimodal interactions between histopathology and genomic profiles for survival prediction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 396–406. Springer (2025)
18. 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)
19. Song, A.H., Chen, R.J., Jaume, G., Vaidya, A.J., Baras, A.S., Mahmood, F.: Multimodal prototyping for cancer survival prediction. arXiv preprint arXiv:2407.00224 (2024)
20. Song, A.H., Jaume, G., Williamson, D.F., Lu, M.Y., Vaidya, A., Miller, T.R., Mahmood, F.: Artificial intelligence for digital and computational pathology. Nature Reviews Bioengineering **1**(12), 930–949 (2023)
21. Wang, Z., Zhang, Y., Xu, Y., Imoto, S., Chen, H., Song, J.: Histo-genomic knowledge association for cancer prognosis from histopathology whole slide images. IEEE Transactions on Medical Imaging **44**(5), 2170–2181 (2025)
22. Xu, Y., Chen, H.: Multimodal optimal transport-based co-attention transformer with global structure consistency for survival prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 21241–21251 (2023)
23. Zhang, C., Qu, L., Liu, X., Song, Z.: Me-mamba: Multi-expert mamba with efficient knowledge capture and fusion for multimodal survival analysis. Computerized Medical Imaging and Graphics **129**, 102733 (2026)
24. 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)
25. Zheng, T., Jiang, K., Xiao, Y., Zhao, S., Yao, H.: M3amba: Memory mamba is all you need for whole slide image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15601–15610 (2025)
26. Zheng, T., Yao, H., Jiang, K., Xiao, Y., Zhao, S.: Gmmamba: Group masking mamba for whole slide image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9935–9944 (2025)
27. Zheng, T., Yao, H., Zhao, S., Jiang, K., Xiao, Y.: Graphmamba: Whole slide image classification meets graph-driven selective state space model. Pattern Recognition **167**, 111768 (2025)

28. Zhong, J., Zhang, M., Liu, L., Liu, T., Jiang, J., Piao, Y., Xu, R., Tian, F., Sun, W., Bi, H., et al.: Diffusion-based cross-staining feature transformation for whole slide image analysis: From h&e to ihc representation learning. Medical Image Analysis p. 104138 (2026)
29. Zhou, H., Zhou, F., Chen, H.: Cohort-individual cooperative learning for multimodal cancer survival analysis. IEEE Transactions on Medical Imaging **44**(2), 656–667 (2024)
30. Zhou, J., Tang, J., Zuo, Y., Wan, P., Zhang, D., Shao, W.: Robust multimodal survival prediction with conditional latent differentiation variational autoencoder. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10384–10393 (2025)