SurvMamba: State Space Model with Multi-grained Multi-modal Interaction for Survival Prediction

Ying Chen[1], Jiajing Xie[3], Yuxiang Lin[2], Yuhang Song[1], Chen Zhang[2], Wenxian Yang[4], Rongshan Yu[1,2*]
[1]*School of Informatics, Xiamen University, Xiamen, China*,
[2]*National Institute for Data Science in Health and Medicine, Xiamen University, Xiamen, China*,
[3]*College of Computer and Information Sciences , Fujian Acriculture And Forestry University, Fuzhou, China*
[4]*Aginome Scientific, Xiamen, China*
[*] *Corresponding Author (Email: rsyu@xmu.edu.cn)*

2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 979-8-3315-1557-7/25/$31.00 ©2025 IEEE | DOI: 10.1109/BIBM66473.2025.11356727

***Abstract*—Multi-modal learning that combines pathological images with genomic data has significantly enhanced the accuracy of survival prediction. Nevertheless, existing methods have not fully utilized the inherent hierarchical structure within both whole slide images (WSIs) and transcriptomic data, from which better intra-modal representations and inter-modal integration could be derived. Moreover, many existing studies attempt to improve multi-modal representations through attention mechanisms, which inevitably lead to high complexity when processing high-dimensional WSIs and transcriptomic data. Recently, a structured state space model named Mamba emerged as a promising approach for its superior performance in modeling long sequences with low complexity. In this study, we propose Mamba with multi-grained multi-modal interaction (SurvMamba) for survival prediction. SurvMamba is implemented with a Hierarchical Interaction Mamba (HIM) module that facilitates efficient intra-modal interactions at different granularities, thereby capturing more detailed local features as well as rich global representations. In addition, an Interaction Fusion Mamba (IFM) module is used for cascaded inter-modal interactive fusion, yielding more comprehensive features for survival prediction. Comprehensive evaluations on five TCGA datasets demonstrate that SurvMamba outperforms other existing methods in terms of performance and computational cost. Our code is available at https://github.com/CYing18/SurvMamba.**

***Index Terms*—Multi-modal Learning, Survival Prediction, Multi-grained, Mamba**

I. Introduction

Survival prediction evaluates patients' mortality risks, thereby enhancing the clinical decision-making process related to diagnosis and treatment planning [1]. For cancer patients, pathological images and genomic profiles provide critical and interconnected information for patient stratification and survival analysis [2]. For example, pathological images detail the tumor microenvironment, capturing the diversity of cancer cells and immune interactions [3]. At the same time, genomic profiles provide critical insights into cancer cell states and immune system factors that affect dynamic prognostic outcomes [4]. Therefore, integrating pathological images and genomic data through multi-modal learning holds considerable promise to enhance the precision of survival predictions.

This work is supported part by the Wang Deyao Outstanding Graduate Scholarship Program (T1062406) of Xiamen University and the Xiamen University–Junlian Life collaborative project "Center for Disease Digitalization and Financial Innovation" (20243160C0023)

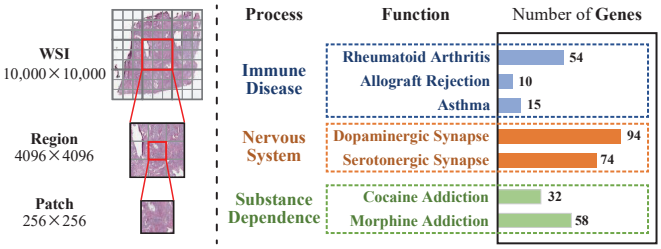

Fig. 1: Hierarchical structure of WSI and transcriptomic data.

Extensive efforts have been made in multi-modal survival analysis with histological whole slide images (WSIs) and transcriptomic data [5], [6]. Among them, Multiple Instance Learning (MIL) [5]–[8] has emerged as an effective method to process the high-dimensional WSIs and transcriptomic data. In these MIL-based methods, each WSI is represented as a "bag" with numerous patches as instances, while transcriptomic data is organized into a "bag" with functional genomes (*i.e.*, genomic groups) as instances. Subsequently, fusion of histological and genomic features [9] is conducted to predict survival outcomes.

Despite the complex, high-dimensional nature of WSIs and transcriptomic data, they exhibit significant inherent hierarchical structures. These structures stem from their fundamental biological functions and the pathology associated with diseases, as illustrated in Figure 1. WSI reveals tissue organization at region-level and provides detailed cellular insights at patch-level. Genes can be categorized according to their genomic function, which can be further subdivided based on biological process. However, existing methods do not fully leverage these hierarchical structures in both WSIs and transcriptomic data [10], which may lead to the following issues. The first issue regards to insufficient global representation. There are multi-level prognostic insights reflected in the hierarchical information. For example, fine-grained patch-level pathological images detail cell densities [11], while coarse-grained region-level images reveal tissue features and tumor-immune interactions [12]. Fine-grained function-level transcriptomics reveals the specific functionalities of gene sets [13], while coarse-grained process-level data focus on identifying those fundamental macroscopic biological processes [14]. Therefore, important global features for survival prediction could

Authorized licensed use limited to: Columbia University Libraries. Downloaded on July 14,2026 at 08:49:32 UTC from IEEE Xplore. Restrictions apply.

potentially be lost if a method focuses on fine-grained information only. Furthermore, there is limited cross-modality communications. Both WSIs and transcriptomic data contain a multitude of cross-modality communications at different hierarchical levels, including patch-to-function and region-to-process relationships [15], which cannot be easily revealed if cross-modality communications are only established at fine-grained level, leading to deficient inter-modal integration.

Although the integration of hierarchical structures into multi-modal survival prediction frameworks holds promise for yielding more comprehensive prognostic insights, it poses significant computational challenges, especially when implemented using attention mechanisms, given their high computational complexity. On the other hand, the Selective Structured State Space Model (SSM), known as Mamba [16], has demonstrated remarkable efficiency in long sequence modeling. Mamba effectively captures long-range dependencies and improves training and inference efficiency through a selection mechanism and a hardware-aware algorithm, hence providing an alternative to deal with high-dimensional data.

In this paper, we propose a Mamba-based survival prediction method with multi-grained multi-modal interaction, termed **SurvMamba**, which extracts multi-modal multi-grained information from hierarchical structure and facilitates efficient intra-modal and inter-modal interactions. Specifically, we design a novel Hierarchical Interaction Mamba (**HIM**) to efficiently capture the hierarchical characteristics of WSIs and transcriptomic data. HIM extracts coarse-grained features through the aggregation of fine-grained instances and enables efficient bidirectional interactions for multi-grained instances. In this way, SurvMamba can extract enhanced local information from numerous fine-grained instances and global information from coarse-grained instances, resulting in more comprehensive intra-modal information to predict patient survival outcomes. Furthermore, we propose an Interaction Fusion Mamba (**IFM**) to facilitate interactions between histological and genomic features across different granularities, thereby providing refined intra-modal representations at both fine and coarse levels. Finally, these multi-grained features are adaptively integrated to formulate the final survival prediction. Our main contributions can be summarized as follows:

- This work is the first attempt to introduce the Mamba model into multi-modal survival prediction, effectively processing high-dimensional WSIs and transcriptomic data with promising performance.
- We propose a Hierarchical Interaction Mamba module to efficiently encode more comprehensive intra-modal representations at both fine-grained and coarse-grained levels from WSI and transcriptomic data.
- We introduce an Interaction Fusion Mamba module, designed to facilitate interaction and integration of histological and genomic features across various levels, thereby capturing multi-modal features from diverse perspectives.
- Extensive experiments are conducted on five public TCGA datasets, and results show that SurvMamba outperforms a variety of state-of-the-art methods with a smaller computational cost.

II. Methodology

A. Overview and Problem Formulation

As illustrated in Figure 2, we propose a novel state space model with multi-grained multi-modal interaction named **SurvMamba** for survival prediction. Clinical data for each patient is encapsulated in a quadruple $X_i = (I_i, G_i, c_i, t_i)$, where I_i represent the set of WSIs, G_i refers to the set of transcriptomics, $c_i \in \{0, 1\}$ is the censoring status and $t_i \in \mathbb{R}^+$ denotes overall survival time (in months). Our objective is to employ WSI I_i and transcriptomic data G_i to estimate hazard functions $f^i_{hazard}(t)$, which represent the probability of a death event occurring in a brief interval following time t for the i-th patient.

To capture more comprehensive multi-modal representations, we exploit the inherent hierarchical structure of WSI and transcriptomic data to extract multi-grained information, subsequently enabling efficient interaction and fusion among them through Mamba-based modules. We adopt the MIL framework to learn pathological and genomic representations, formulating each WSI and transcriptomics as a "bag". We extract features of fine-grained instances (Patch or Function) from the Pathology Encoder and the Genomics Encoder in groups. Then, with bidirectional Mamba in dual-level MIL framework, the Hierarchical Interaction Mamba (**HIM**) module aggregates fine-grained instances into coarse-grained instances (Region or Process) and enables efficient intra-modal interactions at different granularities. Further, the Interaction Fusion Mamba (**IFM**) module facilitates multi-grained interactions between histological and genomic features to derive fine-grained and coarse-grained inter-modal representations. Finally, multi-grained multi-modal features will be adaptively fused to predict a hazard function and get the survival risk scores.

B. Hierarchical Multi-modal Representations

WSIs and transcriptomics exhibit hierarchical structure with multi-level prognostic insights. In this study, we harness this hierarchical nature by representing WSIs and transcriptomics through a three-layer structure. Specifically, WSI is formulated with structures of WSI-, region-, and patch-level, while transcriptomic data are structured across gene-, function-, and process-level, respectively.

For an original WSI I, we split it into M non-overlapping regions R with size of $4,096 \times 4,096$, and each region is further split into N non-overlapping patches P with size of 256×256, where $I = \{R_1, R_2, \ldots, R_M\}$, and $R_m = \{P_{m1}, P_{m2}, \ldots, P_{mN}\}, 1 \leq m \leq M$. We extract the patch-level features with a frozen pre-trained Pathology Encoder $f_I(\cdot)$, resulting in patch-level token sequence T^I_m in R_m, $T^I_m = \{f_I(P_{m1}), f_I(P_{m2}), \ldots, f_I(P_{mN})\}$.

Given a set of transcriptomics measurements G, it can be mapped into J different biological processes S. The j-th process S_j contains K_j genomic functions F, where $G = \{S_1, S_2, \ldots, S_J\}$, and $S_j = \{F_{j1}, F_{j2}, \ldots, F_{jK_j}\}, 1 \leq j \leq J$.

Authorized licensed use limited to: Columbia University Libraries. Downloaded on July 14,2026 at 08:49:32 UTC from IEEE Xplore. Restrictions apply.

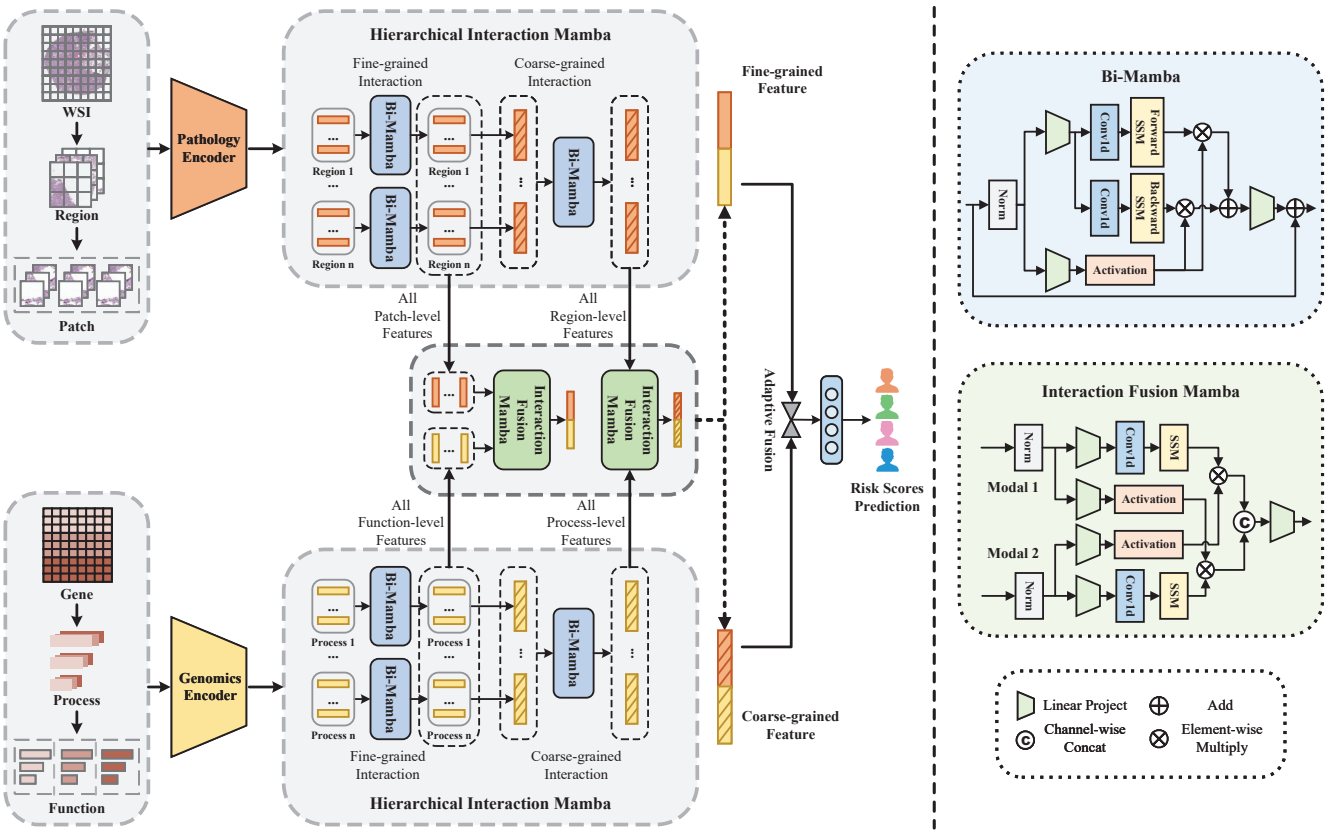

Fig. 2: Overview of SurvMamba architecture. WSIs and transcriptomics are illustrated in a three-layer structure, comprising WSI/Region/Patch for WSIs and Gene/Process/Function for transcriptomics, respectively.

The number of genomic functions contained within each process varies. We encode genomic functions with Genomics Encoder $f_G(\cdot)$ which contains multilayer perceptrons (MLPs) with learnable weights, resulting in function-level token sequence T_j^G in S_j, $T_j^G = \{f_G(F_{j1}), f_G(F_{j2}), \ldots, f_G(F_{jK_j})\}$.

C. Hierarchical Interaction Mamba

Structural WSI and transcriptomic data exhibit certain dependencies within local or global features. To capture these dependencies, we develop the HIM module. Inspired by [17], this module is designed to learn intra-model features by integrating a bidirectional Mamba (Bi-Mamba) mechanism within a dual-level MIL framework. The specific structure of Bi-Mamba is shown in the top-right corner of Figure 2. Our approach aims to model correlations across features of varying granularities effectively. For the first level MIL, the HIM module regards fine-grained (*i.e.*, patch- and function-level) features T_i^I and T_j^G as instances, using Bi-Mamba with shared parameters between groups (*i.e.*, region or process) to model fine-grained histological and genomic long sequences, resulting in enhanced fine-grained features $T_m^{I\prime}$ and $T_j^{G\prime}$:

$$\begin{aligned} T_m^{I\prime} &= \textbf{Bi-Mamba}(T_m^I) \\ T_j^{G\prime} &= \textbf{Bi-Mamba}(T_j^G). \end{aligned} \tag{1}$$

Subsequently, with the second level MIL, features $T_m^{I\prime}$ and $T_j^{G\prime}$ are aggregated into coarse-grained (*i.e.*, region- and process-level) features by average pooling, and Bi-Mamba is then re-applied to facilitate interactions among them to get improved coarse-grained features T^I and T^G. Through this dual-level deep interaction mechanism for each modality, the HIM ensures a thorough and efficient integration of local fine-grained and global coarse-grained features, obtaining enhanced and comprehensive intra-modal histological and genomic features.

$$\begin{aligned} T^I &= \textbf{Bi-Mamba}([\text{Pool}(T_1^{I\prime}), \text{Pool}(T_2^{I\prime}), \ldots, \text{Pool}(T_M^{I\prime})]) \\ T^G &= \textbf{Bi-Mamba}([\text{Pool}(T_1^{G\prime}), \text{Pool}(T_2^{G\prime}), \ldots, \text{Pool}(T_J^{G\prime})]). \end{aligned} \tag{2}$$

D. Interaction Fusion Mamba

To facilitate cross-modal feature interaction and fusion at different granularities, we introduce an IFM module, which is shown in the middle-right part of Figure 2. In IFM, we project features from two modalities and employ gating mechanisms to encourage complementary feature learning from each other while suppressing redundant features. After that, multi-modal features are concentrated to form a fused representation. Cross-modality communications are established at fine-grained and coarse-grained level via cascaded IFM. Through this block, we can obtain fine-grained fused features H_f from features

Authorized licensed use limited to: Columbia University Libraries. Downloaded on July 14,2026 at 08:49:32 UTC from IEEE Xplore. Restrictions apply.

TABLE I: c-index (mean ± std) over five TCGA datasets. G. and H. refer to genomic modality (transcriptomics) and histological modality (WSI), respectively. The best results and the second-best results are highlighted in **bold** and in underline.

Model	G.	H.	BRCA	BLCA	COADREAD	UCEC	LUAD	Overall
SNN	✓		0.606±0.011	0.610±0.038	0.617±0.025	0.610±0.032	0.589±0.031	0.606
SNNTrans	✓		0.621±0.032	0.611±0.010	0.635±0.044	0.592±0.017	0.602±0.044	0.612
ABMIL		✓	0.613±0.033	0.588±0.033	0.624±0.050	0.618±0.014	0.604±0.043	0.609
CLAM-SB		✓	0.605±0.062	0.602±0.031	0.598±0.036	0.576±0.043	0.586±0.033	0.593
CLAM-MB		✓	0.611±0.041	0.609±0.010	0.611±0.036	0.589±0.023	0.612±0.022	0.606
TransMIL		✓	0.628±0.015	0.604±0.045	0.627±0.425	0.601±0.030	0.626±0.030	0.617
R^2T-MIL		✓	0.641±0.006	0.611±0.041	0.604±0.029	0.618±0.030	0.624±0.029	0.620
CLAM-MB (Cat)	✓	✓	0.628±0.047	0.619±0.032	0.614±0.021	0.601±0.031	0.610±0.032	0.614
CLAM-MB (KP)	✓	✓	0.655±0.045	0.633±0.027	0.651±0.053	0.637±0.021	0.629±0.061	0.641
TransMIL (Cat)	✓	✓	0.651±0.039	0.631±0.031	0.636±0.026	0.622±0.043	0.641±0.033	0.636
TransMIL (KP)	✓	✓	0.671±0.021	0.656±0.038	0.661±0.034	0.649±0.036	0.652±0.054	0.658
CMTA	✓	✓	0.687±0.015	0.689±0.035	0.663±0.040	0.685±0.006	0.672±0.025	0.679
MCAT	✓	✓	0.670±0.032	0.669±0.026	0.667±0.025	0.660±0.032	0.682±0.042	0.670
MOTCat	✓	✓	0.692±0.036	0.688±0.029	0.669±0.042	0.692±0.024	0.687±0.046	0.686
SurvPath	✓	✓	0.713±0.025	0.707±0.014	0.683±0.022	0.720±0.026	0.684±0.025	0.701
SurvMamba	✓	✓	**0.737±0.014**	**0.720±0.027**	**0.697±0.018**	**0.731±0.012**	**0.702±0.020**	**0.717**

$T^{I'}$ and $T^{G'}$, and get coarse-grained fused features H_c from features T^I and T^G, as follows:

$$\begin{aligned} H_f &= \mathbf{IFM}([T^{I'}, T^{G'}]) \\ H_c &= \mathbf{IFM}([T^I, T^G]). \end{aligned} \tag{3}$$

E. Survival Prediction

Recognizing the varying significance of features at different granularities for survival prediction, we employ an adaptive fusion strategy with learnable weight denoted as α to fuse features across granular levels. We initialize α at 0.5 and include it in the optimizer parameters, adjusting it during optimization to minimize the loss function. Subsequently, we derive the final feature set for prognostication with H_f and H_c, as follows:

$$H = \alpha H_f + (1 - \alpha) H_c. \tag{4}$$

Survival prediction estimates the risk probability of an outcome event before a specific time. For the final multi-modal feature H^i of i-th patient, we use NLL loss [18] as the loss function for optimizing survival prediction, following [6].

III. Experiments

A. Datasets and Settings

Datasets. To demonstrate the performance of our proposed method, we conducted a series of experiments using five public cancer datasets from The Cancer Genome Atlas (TCGA, http://www.cancer.gov/tcga), which include paired diagnostic WSIs and transcriptomic data alongside verified survival outcomes. The datasets encompass Breast Invasive Carcinoma (BRCA) with 869 cases, Bladder Urothelial Carcinoma (BLCA) with 359 cases, Colon and Rectum Adenocarcinoma (COADREAD) with 296 cases, Uterine Corpus Endometrial Carcinoma (UCEC) with 480 cases, and Lung Adenocarcinoma (LUAD) with 453 cases. Regarding transcriptomic data, we identify 352 unique genomic functions and 42 biological processes, as cataloged in the Kyoto Encyclopedia of Genes and Genomes database (KEGG, https://www.genome.jp/kegg/).

Implementation. For WSI preprocessing, we segmented each WSI into regions of 4096×4096 pixels at a magnification level of $20\times$, subsequently subdividing these regions into patches of 256×256 pixels. The RAdam optimizer was utilized to facilitate model optimization, with a batch size set to 1, a learning rate of 2×10^{-4}, and a weight decay parameter of 5×10^{-3}. The Pathology Encoder [19] generates 768-dimensional embeddings that are projected to 512 dimensions. The Genomics Encoder comprises a two-layer feed-forward network designed to produce genomic tokens with 512 dimensions. We set SSM dimension to 16. To enhance the robustness of model training, 5-fold cross-validation was applied on all models. The models are evaluated using the concordance index (c-index), where a higher value indicates better performance. All computational experiments were conducted on an NVIDIA-A800 GPU.

B. Comparisons with the State-of-the-Art

Baselines. (1) Unimodal baselines. For transcriptomic data, we specifically implemented SNN [20] and SNNTrans [20], [21]. For histology, we compared the MIL methods ABMIL [22], CLAM [23], TransMIL [24], R^2T-MIL [25]. (2) Multi-modal baselines. We compared SOTA methods for multi-modal survival outcome prediction with the previous set-based network architectures (CLAM, TransMIL) with concatenation (Cat) and Kronecker product (KP), two common late fusion mechanisms to integrate bag-level WSI features and genomic features as multi-modal baselines. We also compared CMTA [26], MCAT [27], MOTCat [6] and SurvPath [5] four SOTA methods for multi-modal survival outcome prediction. Table I shows the results of all methods on 5 TCGA datasets.

Unimodal v.s. Multi-modal. In comparison with methods for genomic data, SNN and SNNTrans, SurvMamba outperformed

Authorized licensed use limited to: Columbia University Libraries. Downloaded on July 14,2026 at 08:49:32 UTC from IEEE Xplore. Restrictions apply.

TABLE II: We compare the effects of different components on the performance (c-index) of SurvMamba.

Model	HIM		IFM		BRCA	BLCA	COADREAD	UCEC	LUAD	Overall
	fine	coarse	fine	coarse						
A					0.688±0.023	0.667±0.031	0.660±0.031	0.686±0.015	0.668±0.047	0.674
B	✓				0.708±0.011	0.683±0.065	0.666±0.022	0.717±0.023	0.674±0.309	0.689
C	✓	✓			0.707±0.007	0.691±0.030	0.681±0.030	0.719±0.020	0.675±0.033	0.695
D	✓	✓	✓		0.711±0.081	0.700±0.034	0.684±0.036	0.721±0.017	0.689±0.016	0.701
E	✓	✓		✓	0.709±0.010	0.696±0.013	0.687±0.030	0.720±0.010	0.679±0.019	0.698
F	✓	✓	✓	✓	0.737±0.014	0.720±0.027	0.697±0.018	0.731±0.012	0.702±0.020	0.717

them on all benchmarks, with overall c-index performance increases of 11.1% and 10.5%, respectively. Against the pathology baselines, SurvMamba improved on all the pathology-based unimodal approaches, with performance improvements in the overall c-index ranging 10.0% to 12.4%, demonstrating the merit of integrating histopathology and genomic features. The comparison results also highlight the benefits of utilizing multi-modality in survival prediction.

Multi-modal SOTA v.s. SurvMamba. SurvMamba outperforms all multi-modal approaches with an overall c-index performance increase ranging from 1.6% to 10.3%. When compared with MIL-based multi-modal methods with different fusion methods, SurvMamba achieved a performance increase in overall c-index ranging from 5.9% to 10.3%, highlighting the effectiveness of the proposed multi-modal integration method. SurvMamba also outperforms other SOTA multi-modal learning methods by a obvious margin, showcasing its outstanding ability for multi-modal learning for prognosis.

Computational Cost Comparison. Existing frameworks for multi-modal survival prediction are based on the attention mechanism to model the intra- and inter-relationships, while our method leverages the state space model. To validate the computational efficiency of our method, we compared SurvMamba with some multi-modal SOTA methods which show the top 5 on the overall c-index of the five TCGA datasets. The computational complexity is evaluated using model parameters (Params) and floating-point operation count (FLOPs). Params evaluates the network's scale, while FLOPs assess the model's complexity. As shown in Figure 3, we report the overall c-index on the TCGA dataset, Params, and FLOPs for different SOTA models. Compared to CMTA, the model size and GPU memory of SurvMamba are reduced by 95.70% and 91.65% with a 3.8% better overall c-index. Compared to the second best method SurvPath, SurvMamba also decreases by 53.3% and 51.6% in model size and GPU memory with a 1.6% better c-index. These results show that our model achieves superior performance compared to the state-of-the-art methods with reduced computational cost.

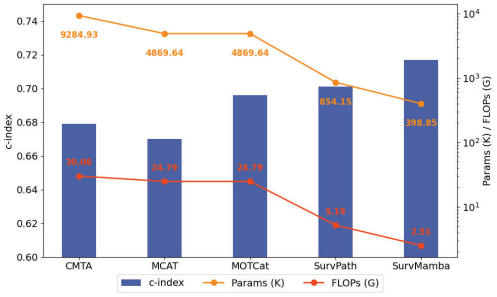

Fig. 3: Computational complexity analysis.

C. Ablation Study

Ablation of components. We conducted ablation studies to validate the effectiveness of the proposed modules. Detailed experimental setups are as follows:

(A) **None**: All fine-grained histological or genomic features are fed as an input vector into the Bi-Mamba block. These unimodal features are then concatenated to predict survival outcomes.

(B) **Fine-grained HIM**: Utilizes shared Bi-Mamba in HIM to refine fine-grained features across specific groups, enhancing these features for prediction without considering coarse-grained information.

(C) **Multi-grained HIM**: Extends Model (B) by incorporating coarse-grained information and integrating multi-grained data for prediction.

(D) **Multi-grained HIM + Fine-grained IFM**: Model (C) with IFM for fine-grained features.

(E) **Multi-grained HIM + Coarse-grained IFM**: Model (C) with IFM for coarse-grained features.

(F) **Multi-grained HIM + Multi-grained IFM**: Model (C) with IFM for multi-grained features (*i.e.*, SurvMamba).

Table II illustrates that Model B, by grouping and processing fine-grained features rather than directly learning from extensive sequences of these features like Model A, more effectively identifies nuanced local relationships. This methodological shift results in a c-index increase from 0.674 to 0.689. The introduction of coarse-grained information, representing broader characteristics in Model C further enhances the c-index to 0.695. These outcomes highlight the advantage of a hierarchical approach in harnessing multi-grained information for prognostic purposes. Improving upon Model C, the application of IFM to integrate and interact with fine- or coarse-grained features resulted in a c-index of 0.701 and 0.689, respectively. This demonstrates the IFM module's effectiveness in integrating multi-modal features. Our proposed SurvMamba model, facilitating both intra-modal and inter-modal interactions and integrations across different granularity levels, delivers a notable c-index of 0.717. Results of the ablation study illustrate the benefits of integrating fine and coarse-grained features and the promising SSM architecture

Authorized licensed use limited to: Columbia University Libraries. Downloaded on July 14,2026 at 08:49:32 UTC from IEEE Xplore. Restrictions apply.

TABLE III: We ablate Mamba modules by replacing them with Transformer models.

Model	BRCA	BLCA	COADREAD	UCEC	LUAD	Time	Flops
Transformer	0.721±0.016	0.702±0.046	0.696±0.021	0.705±0.052	0.679±0.035	0.78s	9.73G
Mamba	0.737±0.014	0.720±0.027	0.697±0.018	0.731±0.012	0.702±0.020	0.22s	2.51G

for survival outcome prediction.

Ablation of Mamba. We conducted an ablation study by replacing Mamba with Transformer to demonstrate the effectiveness of using the Mamba model for WSIs and transcriptomic data. We measured both the inference runtime and FLOPs for the Transformer-based and Mamba-based approaches. As shown in Table III, the Mamba-based method demonstrated greater performance with reduced computational costs.

IV. Conclusion

In this paper, we proposed a novel state space model with multi-grained multi-modal interaction, termed SurvMamba, for survival prediction from WSIs and transcriptomic data. SurvMamba utilizes multi-grained information from hierarchical structures in WSIs and transcriptomic data. We introduced a HIM module that facilitates efficient interactions between intra-modal features at different granularity levels, thereby enriching the unimodal representations. Furthermore, we introduced an IFM module to integrate inter-modal features across various levels, capturing more comprehensive multi-modal features for survival analysis. The experimental results demonstrate SurvMamba's superiority in both performance and computational efficiency, underlining its potential for clinical utilization, such as informing diagnostic and treatment choices for cancer patients.

References

[1] K. A. Tran, O. Kondrashova, A. Bradley, E. D. Williams, J. V. Pearson, and N. Waddell, "Deep learning in cancer diagnosis, prognosis and treatment selection," *Genome Medicine*, vol. 13, pp. 1–17, 2021.

[2] W. Shao, Z. Han, J. Cheng, L. Cheng, T. Wang, L. Sun, Z. Lu, J. Zhang, D. Zhang, and K. Huang, "Integrative analysis of pathological images and multi-dimensional genomic data for early-stage cancer prognosis," *IEEE transactions on medical imaging*, vol. 39, no. 1, pp. 99–110, 2019.

[3] C.-P. Gui, Y.-H. Chen, H.-W. Zhao, J.-Z. Cao, T.-J. Liu, S.-W. Xiong, Y.-F. Yu, B. Liao, Y. Cao, J.-Y. Li *et al.*, "Multimodal recurrence scoring system for prediction of clear cell renal cell carcinoma outcome: a discovery and validation study," *The Lancet Digital Health*, vol. 5, no. 8, pp. e515–e524, 2023.

[4] P. Milanez-Almeida, A. J. Martins, R. N. Germain, and J. S. Tsang, "Cancer prognosis with shallow tumor rna sequencing," *Nature medicine*, vol. 26, no. 2, pp. 188–192, 2020.

[5] G. Jaume, A. Vaidya, R. Chen, D. Williamson, P. Liang, and F. Mahmood, "Modeling dense multimodal interactions between biological pathways and histology for survival prediction," *arXiv preprint arXiv:2304.06819*, 2023.

[6] Y. Xu and H. Chen, "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*, 2023, pp. 21 241–21 251.

[7] L. Qiu, A. Khormali, and K. Liu, "Deep biological pathway informed pathology-genomic multimodal survival prediction," *arXiv preprint arXiv:2301.02383*, 2023.

[8] H. Li, Y. Chen, Y. Chen, R. Yu, W. Yang, L. Wang, B. Ding, and Y. Han, "Generalizable whole slide image classification with fine-grained visual-semantic interaction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 11 398–11 407.

[9] R. J. Chen, M. Y. Lu, D. F. Williamson, T. Y. Chen, J. Lipkova, Z. Noor, M. Shaban, M. Shady, M. Williams, B. Joo *et al.*, "Pan-cancer integrative histology-genomic analysis via multimodal deep learning," *Cancer Cell*, vol. 40, no. 8, pp. 865–878, 2022.

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

[11] D. Zhang, A. Schroeder, H. Yan, H. Yang, J. Hu, M. Y. Lee, K. S. Cho, K. Susztak, G. X. Xu, M. D. Feldman *et al.*, "Inferring super-resolution tissue architecture by integrating spatial transcriptomics with histology," *Nature Biotechnology*, pp. 1–6, 2024.

[12] Y. Fu, A. W. Jung, R. V. Torne, S. Gonzalez, H. Vöhringer, A. Shmatko, L. R. Yates, M. Jimenez-Linan, L. Moore, and M. Gerstung, "Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis," *Nature cancer*, vol. 1, no. 8, pp. 800–810, 2020.

[13] R. Ietswaart, B. M. Gyori, J. A. Bachman, P. K. Sorger, and L. S. Churchman, "Genewalk identifies relevant gene functions for a biological context using network representation learning," *Genome biology*, vol. 22, pp. 1–35, 2021.

[14] X. Zhang, Y. Hu, and D. R. Smith, "Protocol for hsdfinder: Identifying, annotating, categorizing, and visualizing duplicated genes in eukaryotic genomes," *STAR protocols*, vol. 2, no. 3, p. 100619, 2021.

[15] B. J. Smug, K. Szczepaniak, E. P. Rocha, S. Dunin-Horkawicz, and R. J. Mostowy, "Ongoing shuffling of protein fragments diversifies core viral functions linked to interactions with bacterial hosts," *Nature Communications*, vol. 14, no. 1, p. 7460, 2023.

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

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

[18] S. G. Zadeh and M. Schmid, "Bias in cross-entropy-based training of deep survival networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 9, pp. 3126–3137, 2020.

[19] X. Wang, S. Yang, J. Zhang, M. Wang, J. Zhang, J. Huang, W. Yang, and X. Han, "Transpath: Transformer-based self-supervised learning for histopathological image classification," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24*. Springer, 2021, pp. 186–195.

[20] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, "Self-normalizing neural networks," *Advances in neural information processing systems*, vol. 30, 2017.

[21] D. G. Kleinbaum and M. Klein, *Survival analysis a self-learning text*. Springer, 1996.

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

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

[24] Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang, X. Ji *et al.*, "Transmil: Transformer based correlated multiple instance learning for whole slide image classification," *Advances in neural information processing systems*, vol. 34, pp. 2136–2147, 2021.

[25] W. Tang, F. Zhou, S. Huang, X. Zhu, Y. Zhang, and B. Liu, "Feature re-embedding: Towards foundation model-level performance in computational pathology," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 11 343–11 352.

[26] F. Zhou and H. Chen, "Cross-modal translation and alignment for survival analysis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 21 485–21 494.

[27] R. J. Chen, M. Y. Lu, W.-H. Weng, T. Y. Chen, D. F. Williamson, T. Manz, M. Shady, and F. Mahmood, "Multimodal co-attention transformer for survival prediction in gigapixel whole slide images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4015–4025.

Authorized licensed use limited to: Columbia University Libraries. Downloaded on July 14,2026 at 08:49:32 UTC from IEEE Xplore. Restrictions apply.