Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Dynamic Graph and Multi-Stage Mamba Fusion for Multimodal Survival Prediction

Rui Wang

wangrui026@outlook.com

Yun Gu

shanghaijiaotong university

Jie Yang

shanghaijiaotong university

Wei Xu

Tsinghua University

Research Article

Keywords: Multimodal survival prediction, Whole-slide image, Genomics, Dynamic graph, Mamba, Multi-stage fusion

Posted Date: November 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11124211/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Dynamic Graph and Multi-Stage Mamba Fusion for Multimodal Survival Prediction

Rui Wang[1], Yun Gu[2,3], Jie Yang[2*], Wei Xu[1,4*]

[1]Shanghai Qi Zhi Institute, Yunjin Road 701, 40 F, Shanghai, 200232, China.

[2*]Institute of Image Processing and Pattern Recognition, Department of Automation, Shanghai Jiao Tong University, Dongchuan Road 800, Shanghai, 200240, China.

[3]Institute of Medical Robotics, Shanghai Jiao Tong University, Dongchuan Road 800, Shanghai, 200240, China.

[4]Institute for Interdisciplinary Information Sciences, Tsinghua University, Haidian District, Beijing, 100084, China.

Abstract

Multimodal survival prediction using whole-slide images (WSIs) and genomic profiles is important for cancer prognosis; however, integrating gigapixel histopathology with high-dimensional molecular data remains challenging. Existing methods often rely on dense or fixed fusion strategies that may introduce redundant interactions and inadequately capture patient-specific pathology–molecular associations. We propose a sample-adaptive WSI-omics framework that constructs sparse pathway–patch interactions and combines graph-based refinement with selective state-space aggregation for patient-level survival prediction. The framework was evaluated on five TCGA cohorts using patient-level 5-fold cross-validation. Compared with SurvPath, the proposed method improved the average C-index from 0.609 to 0.638, corresponding to an absolute gain of 0.029. The improvement reached significance only in BLCA under the paired t-test, indicating cohort-dependent statistical support. Matched profiling showed lower memory consumption than the Transformer replacement, while case-level visualizations provided model-derived pathway–patch associations. These results support sample-adaptive sparse interaction modeling for WSI-omics survival prediction. External validation, repeated-seed evaluation, and pathologist-guided interpretation remain necessary before clinical application.

Keywords: Multimodal survival prediction, Whole-slide image, Genomics, Dynamic graph, Mamba, Multi-stage fusion

1 Introduction

Survival analysis is used in medical research to evaluate the probability of survival of patients within a specific time frame, which is of great significance for disease prognosis prediction. Traditional survival analysis relied on statistical methods, such as the Cox proportional hazards model[1], which demonstrated robust performance in structured data settings. However, with the advent of deep learning and the growing availability of multimodal data, survival prediction models are increasingly leveraging deep learning to capture complex patterns from diverse sources, including pathological images[2–6], genomic data[7–9], and clinical records[10, 11]. Pathological images, often acquired as gigapixel whole-slide images (WSIs) through high-throughput scanners, offer invaluable spatial and cellular-level insights into disease pathology. These images present challenges due to their immense resolution, necessitating their division into smaller patches for computational processing. Each patch represents a local perspective of the WSI and can be aggregated to characterize the global structure. In this context, multi-instance learning (MIL) [12] has emerged as a powerful paradigm for WSI analysis, treating each WSI as a "bag" and its patches as "instances." By selecting the most survival-relevant instances through feature aggregation, MIL allows for a comprehensive representation of the tumor microenvironment, contributing to precise survival prediction. In parallel, genomic data reveal molecular-level biological characteristics, encompassing gene expression levels, mutation statuses, and copy number variations. These data, typically organized as a matrix with genes as columns and samples as rows, provide a complementary perspective to histopathology images. While the high dimensionality of genomic data poses challenges for computational modeling, dimensionality reduction techniques, such as pathway-based grouping, help extract key survival-related features while enhancing interpretability. Pathways, representing sets of genes involved in specific biological functions or signaling processes, reduce the dimensionality of genomic data from tens of thousands of features to a manageable number, facilitating integration with other data modalities.

Multimodal survival models[13–16] that integrate pathology images and genomic data have shown considerable potential for cancer prognosis. By integrating spatial and morphological tissue features with molecular information, these models can characterize complementary aspects of tumor heterogeneity. Their design nevertheless remains challenging because WSIs and genomic profiles differ substantially in scale, representation, and feature space. Gigapixel WSIs contain thousands of patch-level observations, whereas genomic profiles are typically represented as high-dimensional matrices of molecular measurements. Establishing clinically meaningful correspondences between these heterogeneous representations therefore requires models that capture both local cross-modal interactions and global patient-level dependencies. Existing approaches to multimodal interaction modeling primarily use attention mechanisms or graph-based models. Attention, particularly within Transformer architectures, effectively captures dependencies within and across modalities. Its quadratic complexity, however, can become a computational bottleneck for WSIs containing thousands of patches and genomic profiles with tens of thousands of features. Graph-based models provide an intuitive alternative for representing multimodal interactions, with nodes denoting

features, such as image patches or gene expression measurements, and edges encoding their relationships. However, conventional graph-based methods often construct edges using predefined rules, such as k-nearest neighbors (kNN) or Euclidean distance. Such static rules may not adequately represent the dynamic, patient-specific relationships between modalities. To address these challenges, we propose a task-specific multimodal survival framework that integrates dynamic cross-modal graph construction with Mamba-based sequence modeling. Rather than treating graph learning and Mamba as isolated methodological novelties, we make the following contributions:

- We formulate WSI-omics fusion as sample-adaptive sparse interaction modeling, where pathway tokens query WSI patch tokens and retain only the most relevant Top-k cross-modal edges.
- We combine graph attention and Mamba in a staged pipeline: modality-specific encoding, sparse cross-modal graph refinement, and final long-range aggregation for discrete-time survival prediction.
- We provide a clearer evaluation protocol over five TCGA cohorts, including reproduced baselines, fold-level variability, paired statistical tests against SurvPath, stage-wise ablation, graph-construction ablation, efficiency measurement, and interpretability analysis.

Together, these components constitute a task-specific integration for multimodal survival prediction, rather than a claim of novelty for graph modeling or Mamba in isolation.

2 Related work

Survival analysis predicts the time to a clinical event and is widely used for cancer prognosis and personalized treatment planning [17]. Recent deep learning studies have extended survival prediction beyond clinical variables to imaging-derived and molecular biomarkers, including CT- and MRI-based radiomics [18, 19], histopathological features from WSIs [13], and genomic profiles. Collectively, these studies reflect growing clinical interest in integrating heterogeneous prognostic information across biomedical modalities. This work focuses on WSI-omics survival prediction, for which the central challenge is to jointly model gigapixel pathology images, high-dimensional molecular features, right-censored outcomes, and inter-patient heterogeneity.

2.1 Survival Analysis with State Space Modeling

State space models (SSMs) [20–22] have recently emerged as an effective framework for modeling complex dynamics in sequential data, with advances in natural language processing [23] and time-series forecasting [24]. By modeling long-range dependencies through efficient state transitions, SSMs provide an alternative to attention-based mechanisms for computationally demanding tasks. This property has motivated their application to survival analysis [25–28], particularly for large-scale, high-dimensional, and multimodal data. Structured state space models (S4) [29] achieve linear complexity in sequence length through a state-space formulation that avoids explicit self-attention. Mamba [30] further introduces a selection mechanism that adapts state transitions

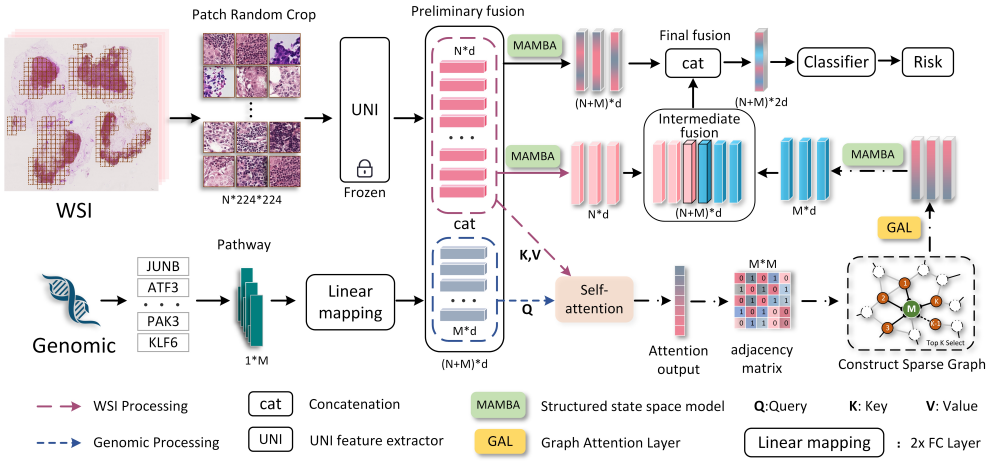


Fig. 1 Overview of the proposed WSI–omics survival prediction framework. For each patient, WSI patches are encoded by UNI into 1,024-dimensional patch embeddings and projected to a 256-dimensional model space. Genomic measurements are grouped into pathway-level tokens and projected into the same feature dimension. Pathway tokens are used as queries and WSI patch tokens as keys/values to construct a sample-adaptive Top-k cross-modal graph. The graph attention module refines the selected pathway–patch interactions, and Mamba blocks aggregate the graph-refined multimodal token sequence. The final prediction head outputs discrete-time survival hazard logits and a patient-level risk score for C-index evaluation.

to the input, enabling efficient content-dependent modeling of long sequences. Recent studies have applied SSMs to both unimodal [25–28] and multimodal [31] survival analysis. Unlike pathology-only Mamba models that primarily aggregate WSI patch sequences, our setting requires explicit interactions between pathway-level genomic tokens and WSI patch tokens before sequence aggregation. We therefore apply Mamba after sample-adaptive graph refinement rather than using it as a standalone WSI encoder.

2.2 Multimodal Survival Analysis

Multimodal survival models integrate complementary information from histopathology, genomics, and clinical variables to characterize the biological and clinical factors associated with patient outcomes. Transformer-based frameworks have been widely adopted to model interactions between modalities. Co-attention Transformers [14, 32, 33] capture cross-modal dependencies for survival prediction. Chen [14] introduced a multimodal co-attention Transformer that reduces the complexity of WSI bag representations while modeling histology–gene interactions. Xu [32] and Li [33] extended this paradigm using optimal transport-based and hierarchical co-attention Transformers, respectively, to address the representation of gigapixel WSIs and interactions within the tumor microenvironment. Other studies [13, 15, 34, 35] have improved the efficiency and interpretability of Transformer-based architectures. For example, Song [13] compresses WSIs into morphological prototypes and transcriptomic data into pathway prototypes, thereby reducing the computational cost of cancer

survival prediction. Jaume [35] employs a memory-efficient multimodal Transformer to fuse histological patch markers with transcriptomic pathway markers and investigate genotype–phenotype interactions. Graph-based approaches [16, 36–38] provide another strategy for survival prediction. For example, [36] presents an attention-based fusion framework that integrates graph-modeled pathology images with gene expression data for survival prediction and tumor classification. These studies demonstrate the potential of Transformer- and graph-based methods for integrating heterogeneous prognostic data. In contrast to MCAT and MOTCat, which use co-attention or optimal-transport-style multimodal alignment, the proposed method first sparsifies WSI–pathway interactions through a patient-specific Top-k graph. Unlike SurvPath, which models dense pathway–histology interactions using Transformer-style fusion, our model explicitly constructs a cross-modal edge set and applies graph attention before Mamba aggregation. Furthermore, unlike TransMIL-style WSI aggregation, our method is explicitly multimodal and evaluates WSI–pathway fusion under the same TCGA split protocol.

2.3 Multimodal Fusion

Effective multimodal fusion [33, 39, 40] should exploit complementary information while accounting for differences in scale, representation, and semantics across modalities. Early fusion combines raw or low-level features at the input stage to learn a shared representation. Although this strategy can capture fine-grained interactions, it may be computationally expensive for high-dimensional inputs such as WSIs and genomic matrices. For example, G. Jaume [35] uses a single-stage fusion strategy in which a Transformer integrates the two modalities at an early stage. Intermediate fusion combines modality-specific features after partial processing by separate encoders, balancing modality-specific representation learning with cross-modal integration. Late fusion [34, 41] aggregates predictions or high-level features from independently trained models. Although modular, this strategy may overlook fine-grained cross-modal relationships.

3 Methods

3.1 Problem Formulation

For each patient, the input consists of WSI patch features and pathway-level genomic features. Let $H = \{h_n\}_{n=1}^{N}$ denote the WSI patch-token sequence and $G = \{g_m\}_{m=1}^{M}$ denote the genomic pathway-token sequence after modality-specific encoders and projection into a shared d-dimensional feature space. The goal is to predict discrete-time survival hazards over J time intervals while properly handling right-censored observations. The central modeling question is how to identify informative WSI–pathway interactions without applying dense attention over all possible token pairs.

The proposed pipeline comprises three stages. First, WSI and pathway tokens are encoded separately to preserve modality-specific information. Second, pathway tokens query WSI tokens to construct a sparse, sample-adaptive cross-modal graph, after which graph attention refines the selected interactions. Third, Mamba blocks

aggregate the early-fused and graph-refined tokens into a patient-level representation for survival prediction. Unlike direct concatenation or dense Transformer fusion, this design selects a patient-specific cross-modal interaction structure before long-range sequence aggregation.

3.2 Preliminaries: Selective Structured State Space Model

WSI-omics survival prediction involves long, heterogeneous token sequences. A single WSI may contain thousands of patch embeddings, whereas omics features are organized as pathway-level molecular tokens. Applying dense self-attention to all pathology and pathway tokens requires pairwise interactions and scales quadratically with sequence length. In addition to its computational cost, dense attention may introduce redundant interactions because only a subset of WSI regions and molecular pathways is expected to be prognostically relevant for an individual patient.

To address this issue, we use Mamba as an efficient long-sequence aggregation module after sparse graph-based cross-modal refinement. Its selective state-space formulation supports content-dependent information propagation along long token sequences with linear complexity in sequence length under the selective scan implementation. In our framework, cross-modal attention first identifies sparse pathway–patch interactions, after which Mamba aggregates the graph-refined multimodal tokens into a patient-level representation. This design exploits the computational properties of Mamba without assuming its universal superiority over Transformers.

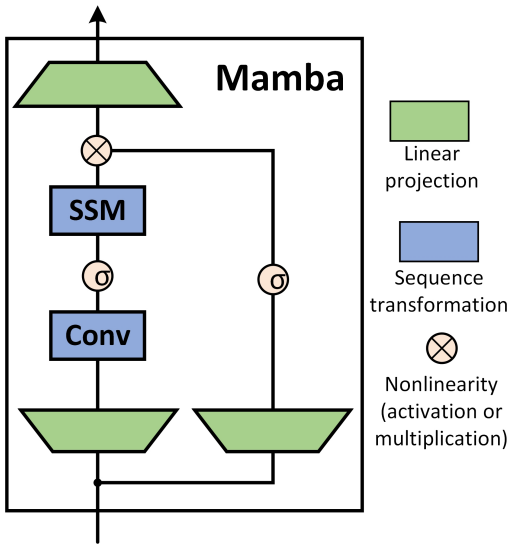


Fig. 2 Mamba block used for multimodal sequence aggregation. It propagates context along graph-refined WSI–pathway tokens with selective state-space modeling, avoiding dense pairwise self-attention.

A continuous state-space model maps an input signal $x(t)$ to an output signal $y(t)$ through a latent state $h(t)$:

$$\begin{aligned} h'(t) &= Ah(t) + Bx(t), \\ y(t) &= Ch(t), \end{aligned} \tag{1}$$

where A, B, and C denote the state transition, input projection, and output projection matrices, respectively. After discretization with step size Δ, the recurrence can be written as

$$\begin{aligned} h_t &= \bar{A}h_{t-1} + \bar{B}x_t, \\ y_t &= Ch_t. \end{aligned} \tag{2}$$

Unlike conventional linear time-invariant state-space models, Mamba introduces input-dependent selective parameters, typically through token-conditioned Δ, B, and C. Consequently, the model can modulate how information is written to, maintained in, and read from the latent state according to the current token content. We use this property to aggregate the refined WSI–pathway token sequence after graph-based selection of prognostically relevant cross-modal interactions.

3.3 Dynamic Graph Construction for Modal Interaction

The graph module identifies the WSI regions that interact with each genomic pathway token for patient-specific survival prediction. Rather than using a fixed graph or a global nearest-neighbor rule, we infer a sparse cross-modal graph from attention scores between pathway and WSI patch tokens. Let $G = \{g_m\}_{m=1}^{M}$ denote the pathway-level genomic tokens and $H = \{h_n\}_{n=1}^{N}$ denote the WSI patch tokens. The pathway tokens serve as queries, and the WSI patch tokens serve as keys and values:

$$Q = GW_Q, \quad K = HW_K, \quad V = HW_V, \tag{3}$$

where W_Q, W_K, and W_V are learnable projection matrices. The cross-modal attention matrix is computed as

$$A = \mathrm{Softmax}\left(\frac{QK^\top}{\sqrt{d}}\right), \tag{4}$$

where A_{mn} measures the relevance between pathway token g_m and WSI patch token h_n.

For each pathway token, we retain only the Top-k WSI patch tokens with the highest attention scores:

$$\begin{aligned} S_m &= \mathrm{TopK}\left(A_{m1}, A_{m2}, \ldots, A_{mN}; k\right), \\ \mathcal{E} &= \{(g_m, h_n) \mid n \in S_m\}. \end{aligned} \tag{5}$$

The resulting edge set $\mathcal{E}$ defines a patient-specific bipartite graph between genomic pathway tokens and WSI patch tokens. This sparse construction restricts cross-modal fusion to the most relevant pathway–patch interactions, rather than forcing every pathway token to interact with all WSI patches. In the main setting, we use $k = 8$,

and the graph-construction ablation evaluates alternative sparsity levels and graph construction strategies.

3.4 Graph Attention Refinement

Given the sparse pathway–patch graph, we apply graph attention to refine the selected cross-modal interactions. For a pathway node g_m, its neighbor set is defined by the retained WSI patch tokens:

$$\mathcal{N}(g_m) = \{h_n \mid (g_m, h_n) \in \mathcal{E}\}. \tag{6}$$

The refined pathway representation is computed as

$$\tilde{g}_m = \sum_{h_n \in \mathcal{N}(g_m)} \alpha_{mn} W h_n, \tag{7}$$

where W is a learnable transformation matrix and α_{mn} is the graph attention coefficient:

$$\alpha_{mn} = \frac{\exp\left(\text{LeakyReLU}\left(a^\top [W_g g_m \| W_h h_n]\right)\right)}{\sum_{h_r \in \mathcal{N}(g_m)} \exp\left(\text{LeakyReLU}\left(a^\top [W_g g_m \| W_h h_r]\right)\right)}. \tag{8}$$

Here, W_g and W_h project pathway and WSI patch tokens into a shared interaction space, and $\|$ denotes feature concatenation. This graph refinement step updates pathway representations using only their selected pathology neighbors, thereby reducing unnecessary dense pairwise fusion and producing an explicit pathway–patch edge structure that can be further inspected for interpretation.

The graph-refined pathway tokens are then combined with WSI patch tokens and passed into the Mamba aggregation block. In this way, the proposed pipeline separates sparse cross-modal interaction discovery from long-range sequence aggregation: the graph module identifies and refines patient-specific pathology–omics interactions, while Mamba propagates contextual information across the refined multimodal token sequence to obtain the final patient-level representation for survival prediction.

3.5 Multi-stage Multimodal Fusion

The multi-stage fusion module integrates the two modalities through the following three stages.

Stage 1: early token fusion.

Projected WSI and pathway tokens are concatenated and passed through a Mamba block to obtain an early global multimodal representation. This stage allows the model to capture coarse patient-level context before explicit graph refinement.

Stage 2: sparse graph interaction.

The dynamic Top-k graph identifies the WSI tokens most relevant to each pathway token. Graph attention updates the selected tokens and produces graph-refined cross-modal representations. This stage focuses the model on local cross-modal evidence instead of treating all token pairs equally.

Stage 3: final sequence aggregation.

The early fused representation and graph-refined representation are concatenated and passed through final normalization and prediction layers. A final Mamba aggregation block is used when the full multi-stage variant is enabled. Stage-wise ablations remove or simplify these stages to test whether the final architecture improves beyond empirical stacking.

Algorithm 1 Dynamic graph and multi-stage Mamba fusion

Require: WSI patch features x_{hist}, pathway-level omics features x_{omic}
Ensure: Discrete-time survival hazard logits

$H \leftarrow f_{\text{UNI}}(x_{\text{hist}})$
$G \leftarrow f_{\text{omic}}(x_{\text{omic}})$
$E_1 \leftarrow \text{Mamba}_1(\text{Concat}(H, G))$
$A \leftarrow \text{Softmax}\left((GW_Q)(HW_K)^\top/\sqrt{d}\right)$
for each pathway token g_m **do**
    $S_m \leftarrow \text{TopK}(A_{m,:}; k)$
    add edges $\{(g_m, h_n) \mid n \in S_m\}$ to $\mathcal{E}$
end for
$G' \leftarrow f_{\text{GAT}}(G, H, \mathcal{E})$
$E_2 \leftarrow \text{Mamba}_2(\text{Concat}(G', H))$
$F \leftarrow \text{Concat}(E_1, E_2)$
return $\text{Linear}(F)$

3.6 Survival Prediction Objective

We use a standard discrete-time survival formulation consistent with our implementation. The observed time axis is divided into J non-overlapping intervals. For patient i, let y_i denote the observed interval index and let $c_i \in \{0, 1\}$ denote the censoring indicator, where $c_i = 1$ indicates a censored observation and $c_i = 0$ indicates an observed event. The model outputs hazard logits z_{ij}, which are converted into interval-specific hazard probabilities:

$$h_{ij} = \sigma(z_{ij}), \quad j = 1, \ldots, J. \tag{9}$$

Here, h_{ij} represents the conditional probability that patient i experiences the event in interval j given survival before that interval. The discrete survival probability through

interval j is defined as

$$S_{ij} = \prod_{\ell=1}^{j} (1 - h_{i\ell}), \quad S_{i0} = 1. \tag{10}$$

For an uncensored patient whose event occurs in interval y_i, the likelihood contribution is the probability of surviving all previous intervals and experiencing the event in interval y_i:

$$P(T_i = y_i \mid x_i) = S_{i,y_i-1} h_{i,y_i}. \tag{11}$$

For a censored patient observed through interval y_i, the likelihood contribution is the probability of surviving beyond the observed interval:

$$P(T_i > y_i \mid x_i) = S_{i,y_i}. \tag{12}$$

Therefore, the negative log-likelihood is

$$\begin{aligned} \mathcal{L}_{\mathrm{NLL}} = -\sum_i \Big[& (1 - c_i) \big(\log S_{i,y_i-1} + \log h_{i,y_i} \big) \\ & + c_i \log S_{i,y_i} \Big]. \end{aligned} \tag{13}$$

For numerical stability, hazard probabilities are clipped to $[\epsilon, 1 - \epsilon]$ before computing logarithms. In our experiments, the patient-level risk score used for C-index calculation is computed from the predicted hazard trajectory as

$$r_i = \sum_{j=1}^{J} h_{ij}, \tag{14}$$

where a larger value indicates a higher predicted risk. This standard discrete-time negative log-likelihood objective is used to train both the proposed model and its ablation variants.

4 Results

4.1 Dataset and Data Preprocessing

We evaluated the proposed framework on five cohorts from The Cancer Genome Atlas (TCGA) [42]: Breast Invasive Carcinoma (BRCA), Bladder Urothelial Carcinoma (BLCA), Head and Neck Squamous Cell Carcinoma (HNSC), Lung Adenocarcinoma (LUAD), and Stomach Adenocarcinoma (STAD). These cohorts provide matched whole-slide images (WSIs), genomic profiles, and clinical follow-up data across different tumor types, making them suitable for evaluating multimodal survival prediction.

Clinical annotations and genomic profiles were obtained from UCSC Xena [43]. Following previous WSI-omics survival studies [35], we included patients with available clinical outcomes, RNA expression profiles, and corresponding WSI features. Disease-Specific Survival (DSS) was used as the survival endpoint. For histopathology, WSIs were processed with a CLAM-style pipeline [44]. Tissue regions were segmented, and

Table 1 Summary of the five TCGA cohorts used in the DSS survival prediction experiments after matching clinical metadata, hallmarks RNA profiles, patient-level split files, and available WSI feature files. Raw patch embeddings denote the total number of available WSI patch embeddings extracted from CLAM h5 feature files.

Cohort	Patients	WSI feature files	Raw patch embeddings	Train/Val patients per fold
BRCA	868	928	40,275,113	686–702 / 166–182
BLCA	359	423	28,146,375	284–289 / 70–75
HNSC	392	412	18,288,094	266–336 / 56–126
LUAD	431	481	23,661,943	343–346 / 85–88
STAD	318	342	16,469,355	217–276 / 42–101
Total	2368	2586	126,840,880	–

patch-level features were extracted using the UNI foundation model [45]. For each patient, WSI patch embeddings were grouped into a patient-level bag and paired with the corresponding genomic pathway features. All experiments used patient-level five-fold cross-validation with identical split files across methods. To avoid information leakage, genomic features were normalized in a fold-aware manner: the scaler was fitted only on the training partition and then applied to the corresponding validation partition. Detailed cohort statistics are shown in Table 1.

4.2 Evaluation Setting and Metrics

Table 2 summarizes the main implementation settings. All experiments followed a patient-level five-fold cross-validation protocol. The proposed method and all reproduced baselines were evaluated on the same matched patient set using identical split files, clinical endpoint, WSI feature files, and genomic inputs. Performance was measured using the concordance index (C-index) and is reported as the mean ± standard deviation across the five folds. For the main comparison with SurvPath, we report

Table 2 Key implementation settings of the proposed model.

Item	**Setting**
WSI feature extractor	UNI [45]
WSI feature dimension	1,024 → 256
Patient bag size	4,096 patches
Genomic grouping	Hallmark pathways
Graph construction	Dynamic Top-k, $k = 8$
Survival endpoint	DSS
Survival loss	`nll_surv`, $\alpha = 0.5$
Survival bins	4
Training epochs	20
Optimizer	RAdam
Scheduler	Cosine LR schedule
Dropout	0.1

fold-level 95% confidence intervals and conduct paired t-tests and Wilcoxon signed-rank tests using the five fold-level C-index values. Because both models were evaluated on the same patient partitions, paired tests assess the consistency of the observed performance differences across folds.

WSI inputs were represented by pre-extracted UNI features stored in CLAM `h5` files. Each patient-level bag was capped at 4,096 patch embeddings. During training, bags exceeding this limit were randomly sampled without replacement, while shorter bags were zero-padded with a mask. During validation, larger bags were subsampled at evenly spaced intervals to reduce evaluation randomness.

Genomic pathway tokens were constructed from the same Hallmark signature files for SurvPath-style models and the proposed model. To avoid information leakage, genomic normalization was performed within each training fold: normalization parameters were fitted only on the training partition and then applied to the corresponding validation partition. Learning rate and weight decay were selected through a predefined validation procedure within each cohort, without test-fold-specific tuning.

Within the training pipeline, pathway-level genomic features were scaled to the range [-1, 1] using a MinMax scaler fitted exclusively to the training fold; the fitted scaler was then applied to the corresponding validation fold. No additional logarithmic transformation or batch-normalization step was applied at this stage. The main experiments used a batch size of 1, random seed 1, no weighted sampling, one warmup epoch, and an encoder dropout rate of 0.1. The cohort-specific learning rate and weight decay were 0.0005 and 0.01 for BRCA, 0.001 and 0.01 for BLCA, 0.0005 and 0.001 for HNSC, 0.001 and 0.00001 for LUAD, and 0.0005 and 0.01 for STAD, respectively.

All baseline values in the main comparison were obtained by re-implementing or adapting the corresponding methods and evaluating them on the current preprocessed cohorts; no value was copied directly from the original publications. The labels Cat and KP denote concatenation and bilinear/Kronecker-product fusion, respectively. In particular, the reported TransMIL variants used both WSI and pathway-level genomic inputs and therefore should not be interpreted as WSI-only TransMIL results.

4.3 Quantitative Results Analysis

4.3.1 Overall Survival Prediction Performance

We compared the proposed model with representative WSI-based and WSI-omics survival prediction methods across five TCGA cohorts. The comparison was performed under the same Hallmark pathway setting and patient-level five-fold cross-validation protocol. Table 3 reports the mean C-index and standard deviation across folds.

Overall, the proposed model achieved the highest average C-index across the five cohorts. Relative to SurvPath, the average C-index increased from 0.609 to 0.638, corresponding to an absolute gain of 0.029. The largest improvement was observed in BLCA (+0.056), followed by STAD (+0.035), LUAD (+0.030), and HNSC (+0.024). In BRCA, the gain over SurvPath was small (+0.003), indicating a limited additional benefit when the baseline already performed strongly.

The cohort-specific results do not indicate uniform dominance in every setting. In HNSC, TransMIL(KP) achieved a marginally higher C-index than the proposed

Table 3 C-index comparison across five TCGA cohorts using Hallmark pathway features and patient-level five-fold cross-validation. Results are reported as mean ± standard deviation. Bold indicates the best result within each cohort and the best average performance.

Model	BRCA	BLCA	HNSC	LUAD	STAD	Avg.
AB-MIL(Cat) [46]	0.588 ± 0.070	0.518 ± 0.046	0.521 ± 0.065	0.583 ± 0.032	0.589 ± 0.071	0.560
AB-MIL(KP) [46]	0.532 ± 0.089	0.555 ± 0.070	0.559 ± 0.049	0.531 ± 0.067	0.612 ± 0.027	0.558
DeepMISL(Cat) [47]	0.512 ± 0.032	0.574 ± 0.011	0.588 ± 0.091	0.558 ± 0.004	0.597 ± 0.057	0.568
DeepMISL(KP) [47]	0.593 ± 0.051	0.601 ± 0.023	0.571 ± 0.044	0.569 ± 0.076	0.593 ± 0.085	0.584
MOTCat [32]	0.646 ± 0.011	0.601 ± 0.070	**0.598** ± 0.031	0.515 ± 0.086	0.524 ± 0.022	0.577
MCAT [14]	0.545 ± 0.146	0.533 ± 0.043	0.498 ± 0.084	0.608 ± 0.085	0.559 ± 0.080	0.549
TransMIL(Cat) [48]	0.645 ± 0.099	0.605 ± 0.105	0.499 ± 0.066	0.593 ± 0.059	0.577 ± 0.039	0.584
TransMIL(KP) [48]	0.705 ± 0.168	0.579 ± 0.022	0.573 ± 0.050	0.588 ± 0.088	0.509 ± 0.070	0.591
SurvPath [35]	0.722 ± 0.051	0.576 ± 0.086	0.548 ± 0.088	0.604 ± 0.106	0.596 ± 0.030	0.609
OUR	**0.725** ± 0.061	**0.631** ± 0.071	0.572 ± 0.069	**0.634** ± 0.092	**0.631** ± 0.081	**0.638**

Table 4 Fold-level statistical comparison between the proposed model and reproduced SurvPath. CI denotes the 95% confidence interval of the proposed model across five folds.

Cohort	Ours mean ± std	Ours 95% CI	SurvPath mean ± std	Delta	Paired t / Wilcoxon p
BRCA	0.725 ± 0.061	[0.649, 0.800]	0.722 ± 0.052	+0.003	0.931 / 0.813
BLCA	0.631 ± 0.071	[0.543, 0.719]	0.576 ± 0.086	+0.056	0.009 / 0.063
HNSC	0.572 ± 0.070	[0.485, 0.658]	0.548 ± 0.089	+0.024	0.317 / 0.813
LUAD	0.634 ± 0.092	[0.520, 0.748]	0.604 ± 0.106	+0.030	0.535 / 0.813
STAD	0.631 ± 0.081	[0.530, 0.732]	0.596 ± 0.030	+0.035	0.428 / 0.625

model (0.573 versus 0.572). Thus, the main comparison demonstrates improved average performance across cohorts rather than consistent superiority in every individual cohort.

Across model families, simple fusion or attention-based MIL methods, including AB-MIL, MCAT, and TransMIL variants, achieved average C-indices between 0.549 and 0.571. By contrast, models incorporating graph structure or pathway-aware multimodal modeling achieved higher average performance, with SurvPath reaching 0.609 and the proposed model reaching 0.638. These results suggest that explicitly modeling relationships between histological patterns and molecular pathways can benefit survival prediction compared with global feature concatenation or generic attention pooling alone. To further assess the comparison with SurvPath, we conducted paired statistical analyses using the five fold-level C-index values. Table 3 summarizes the mean performance, confidence intervals, and paired-test results.

The strongest fold-level evidence was observed in BLCA, where the paired t-test reached statistical significance ($p = 0.009$). The remaining cohorts showed positive mean differences but did not reach statistical significance under five-fold testing. These results support a cautious interpretation: the proposed model improves the average C-index over SurvPath, with the most robust statistical evidence observed in BLCA. We therefore avoid claiming statistically significant superiority in every cohort.

Table 5 Controlled fusion-component analysis on five TCGA cohorts. C-index is reported as mean ± standard deviation over five folds.

Variant	BRCA	BLCA	HNSC	LUAD	STAD	Avg.
Single-stage	0.679 ± 0.070	0.564 ± 0.063	0.526 ± 0.063	0.580 ± 0.088	0.575 ± 0.068	0.585
Early-fusion only	0.705 ± 0.049	0.555 ± 0.075	**0.539** ± 0.073	**0.602** ± 0.101	**0.587** ± 0.060	**0.598**
Graph-only	0.487 ± 0.088	0.535 ± 0.094	0.482 ± 0.044	0.563 ± 0.118	0.554 ± 0.123	0.524
Two-stage	**0.723** ± 0.049	**0.574** ± 0.056	0.512 ± 0.070	0.594 ± 0.075	0.570 ± 0.067	0.595
Without final aggregation	0.655 ± 0.106	0.561 ± 0.065	0.520 ± 0.056	0.592 ± 0.082	0.547 ± 0.085	0.575
Full multistage variant	0.699 ± 0.044	0.562 ± 0.053	0.509 ± 0.078	0.591 ± 0.086	0.574 ± 0.056	0.587

Fold-level variability remained non-negligible, with standard deviations of 0.061, 0.071, 0.070, 0.092, and 0.081 for BRCA, BLCA, HNSC, LUAD, and STAD, respectively. The largest variability was observed in LUAD rather than HNSC under the current matched evaluation. Because each comparison contains only five folds, these estimates provide limited statistical power and should not be used to attribute instability to a specific biological or institutional source without additional site-aware or repeated-seed analyses.

4.3.2 Fusion Component Analysis

We next analyzed the effect of different fusion designs on survival prediction under a controlled stage-ablation protocol. This experiment characterizes architectural sensitivity and is distinct from the validation-selected main results in Table 3.

The fusion-component analysis shows that graph interaction alone was insufficient for reliable survival prediction. The graph-only variant achieved the lowest average C-index (0.524), indicating that sparse cross-modal edges cannot replace sequence-level representation learning. In contrast, the early-fusion and two-stage variants achieved higher average C-indices of 0.598 and 0.595, respectively. Thus, coarse multimodal alignment and graph-guided interaction modeling both contributed useful information, although their effectiveness depended on how they were combined.

The controlled ablation also shows that adding more fusion stages does not necessarily lead to monotonic improvement. The full multistage variant did not outperform all simplified variants under this protocol. This result is important because it indicates that the advantage of the framework does not come from simply stacking additional fusion modules. Instead, the model benefits from a balance between sequence-level context, sparse graph refinement, and final aggregation.

These findings help clarify the role of graph modeling in the proposed framework. Graph-based interaction should be viewed as a refinement mechanism for selected pathology–omics relationships, rather than as a standalone replacement for multimodal representation learning. This also explains why the final model selection is supported primarily by the main comparison in Table 3 and the graph-construction analysis, while the stage-ablation experiment is used to characterize the behavior of different fusion components.

This fusion-component analysis is distinct from the modality ablation in Table 7. The present experiment studies how multimodal information is fused within the model, whereas the modality ablation evaluates whether histology and genomics provide complementary prognostic signals.

Table 6 Graph-construction and Top-k ablation on the final Ours architecture under matched main-table hyperparameters. The $k = 8$ row reuses the main Ours results because it is the default final-model setting.

Strategy	BRCA	BLCA	HNSC	LUAD	STAD	Avg.
Dyn. Top-k ($k = 1$)	0.671 ± 0.072	0.563 ± 0.071	0.538 ± 0.081	0.591 ± 0.121	0.599 ± 0.095	0.592
Dyn. Top-k ($k = 4$)	0.703 ± 0.049	0.580 ± 0.045	0.561 ± 0.076	0.599 ± 0.101	0.599 ± 0.041	0.608
Dyn. Top-k ($k = 8$)	**0.725** ± 0.061	**0.631** ± 0.071	**0.572** ± 0.070	**0.634** ± 0.092	**0.631** ± 0.081	**0.638**
Dyn. Top-k ($k = 16$)	0.716 ± 0.041	0.557 ± 0.089	0.544 ± 0.050	0.596 ± 0.107	0.587 ± 0.065	0.600
Dyn. Top-k ($k = 32$)	0.704 ± 0.040	0.598 ± 0.047	0.568 ± 0.077	0.599 ± 0.091	0.578 ± 0.070	0.609
Dense graph	0.712 ± 0.076	0.604 ± 0.065	0.556 ± 0.082	0.627 ± 0.130	0.591 ± 0.077	0.618
Random Top-k ($k = 8$)	0.691 ± 0.064	0.630 ± 0.057	0.550 ± 0.072	0.615 ± 0.095	0.588 ± 0.061	0.615
Static KNN ($k = 8$)	0.691 ± 0.089	0.593 ± 0.082	0.553 ± 0.082	0.592 ± 0.112	0.588 ± 0.105	0.604

4.3.3 Graph Construction and Top-k Ablation

We further examined the dependence of performance on the construction of pathology–omics interactions. The final architecture was held fixed while varying the graph-construction strategy and the number of retained neighbors. The default setting uses dynamic Top-k graph construction with $k = 8$.

Table 6 shows that performance depended on both the graph-construction rule and the sparsity level. The default dynamic Top-k setting with $k = 8$ achieved the highest average C-index. Performance decreased markedly when the graph was too sparse ($k = 1$), suggesting that a single retained interaction was insufficient to capture the range of within-patient pathology–omics associations. Increasing the neighborhood size beyond $k = 8$ did not further improve performance; larger neighborhoods and dense graphs may introduce weak or less relevant edges and thereby reduce the benefit of adaptive interaction selection.

Comparisons with random Top-k and static KNN further support patient-specific graph construction. Although random Top-k was competitive in BLCA, its average performance was lower, and static KNN also underperformed the dynamic graph. Thus, the gain cannot be attributed solely to graph sparsification; selecting edges according to sample-specific cross-modal relevance was important for stable performance across cohorts.

4.3.4 Modality Contribution Analysis

We then evaluated the contribution of each input modality by comparing omics-only, histology-only, and multimodal settings. This analysis assesses whether the two modalities provide complementary prognostic information and is distinct from the preceding analysis of individual fusion stages.

Table 7 shows that the relative contributions of histology and genomics varied across cancer types. Omics-only models performed strongly in BLCA, HNSC, and LUAD, whereas histology-only modeling was also competitive in several cohorts. The proposed multimodal model achieved the highest average C-index and clear gains in BRCA and STAD, but did not outperform the strongest single-modality baseline in every cohort.

Table 7 Modality contribution analysis across five TCGA cohorts. C-index is reported as mean ± standard deviation over five folds. The table compares omics-only, histology-only, and multimodal input settings.

Method	Histo	Omics	BRCA	BLCA	HNSC	LUAD	STAD
SNN	No	Yes	0.587 ± 0.103	0.609 ± 0.082	0.492 ± 0.061	0.638 ± 0.073	0.519 ± 0.056
MLP-Omics	No	Yes	0.610 ± 0.167	0.652 ± 0.066	0.601 ± 0.041	0.646 ± 0.050	0.586 ± 0.115
MLP-WSI	Yes	No	0.644 ± 0.130	0.632 ± 0.060	0.582 ± 0.067	0.644 ± 0.048	0.533 ± 0.079
Ours	Yes	Yes	0.725 ± 0.061	0.631 ± 0.071	0.572 ± 0.070	0.634 ± 0.092	0.631 ± 0.081

Table 8 Matched efficiency comparison under the same profiling setting. The table reports model size, training time, inference time, and peak GPU memory.

Model	Params	Train time / epoch	Inference time / patient	Peak GPU memory
Ours	8.84M	22.1 min	15.6 ms	441 MB
Transformer replacement	7.56M	22.7 min	16.8 ms	630 MB
No-Mamba variant	5.98M	20.1 min	14.4 ms	144 MB
TransMIL-style baseline	2.88M	21.8 min	7.7 ms	125 MB

These results indicate that the benefit of combining WSIs and genomic pathways was cohort-dependent rather than guaranteed by the availability of two data sources. One modality may already contain strong prognostic information in some cohorts, whereas their combination provides a clearer advantage in others. The results therefore support adaptive fusion while cautioning against assuming universal modality complementarity across tumor types.

4.3.5 Computational Efficiency

Finally, we evaluated the computational cost of the proposed model under a matched profiling setting. All models were profiled using the same cohort, fold, patch count, pathway setting, batch size, feature dimension, hardware, and precision. As shown in Table 8, the proposed graph-Mamba design did not incur prohibitive computational costs. Compared with the Transformer replacement, our model required less peak GPU memory (441 MB versus 630 MB) and slightly less training and inference time despite having more parameters. The No-Mamba and TransMIL-style variants were lighter, as expected, but did not include the complete pathology–omics interaction modeling of the proposed framework.

These results indicate that the proposed model remained practical under the matched profiling setting. This efficiency analysis provides a cost comparison and is not a substitute for the full five-fold performance evaluation reported in Table 3.

4.3.6 Risk Stratification and Interpretability

We next examined whether the predicted risk scores stratified patients into groups with different survival outcomes. As shown in Figure 3, patients were divided into high- and low-risk groups using the fold-wise median risk score. The proposed model achieved significant survival separation in BRCA, BLCA, LUAD, and STAD, with log-rank p-values of 0.002, 0.001, < 0.001, and 0.039, respectively. Although HNSC did not reach statistical significance ($p = 0.133$), the two curves showed a visible

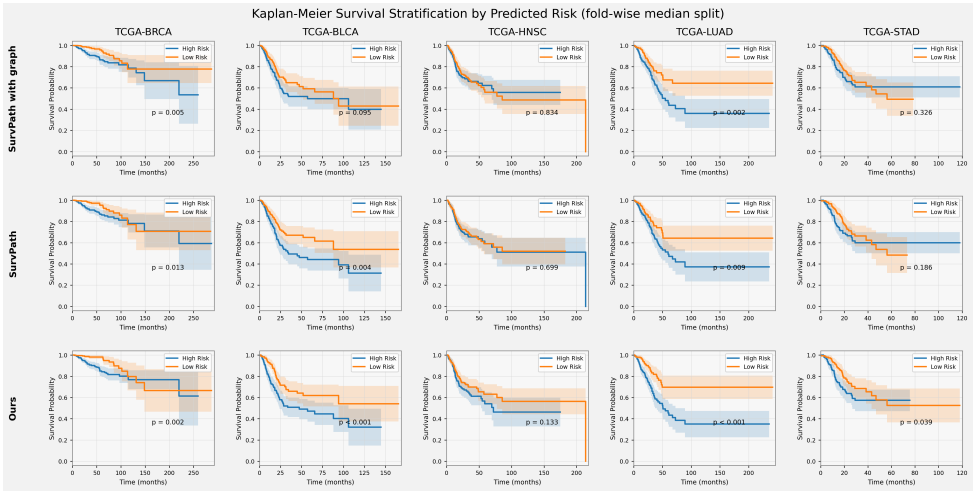


Fig. 3 Kaplan–Meier survival curves for predicted high- and low-risk groups across five TCGA cohorts. The figure uses a 3×5 layout with SurvPath with graph, SurvPath, and Ours as rows, and BRCA, BLCA, HNSC, LUAD, and STAD as columns. Patients were stratified by the fold-wise median predicted risk score and then pooled for visualization. In the Ours row, the two-sided log-rank p-values were 0.002, 0.001, 0.133, < 0.001, and 0.039 for BRCA, BLCA, HNSC, LUAD, and STAD, respectively. Number-at-risk tables are shown below the curves.

separation trend. This analysis complements the C-index results by demonstrating patient-level risk stratification based on the predicted scores. We further visualized the case-level pathology–omics interactions learned by the dynamic graph. In Figure 4, the WSI heatmaps highlight regions with high pathway–patch interaction weights, whereas the enlarged patches show the corresponding local morphology. The bar plots rank the Hallmark pathway tokens by normalized interaction strength. These examples provide patient-specific pathway–patch association patterns in addition to a scalar risk score. The highlighted HNSC and STAD cases exhibited distinct pathway-level interaction profiles associated with spatially localized histological regions. However, these associations represent model-derived explanations rather than validated biological mechanisms. Pathologist review and biological validation are required before mechanistic conclusions can be drawn.

Risk stratification complements continuous C-index evaluation because Kaplan–Meier analysis converts predicted scores into discrete groups. The significant separation in BRCA, BLCA, LUAD, and STAD indicates that the continuous rankings also produced group-level differences under the fold-wise median split, whereas the non-significant HNSC result shows that ranking discrimination does not necessarily translate into significant dichotomized separation. The fold-wise median was used only as an evaluation threshold and should not be interpreted as a transferable clinical cutoff. Similarly, the pathway–patch heatmaps represent relative within-case interaction strengths rather than calibrated biomarkers. Without pathologist annotations or biological validation, the highlighted regions and pathways should be viewed as model-derived hypotheses.

5 Discussion

The proposed framework combines sample-adaptive sparse cross-modal graph construction, graph attention refinement, and Mamba-based sequence aggregation for WSI-omics survival prediction. The design addresses the potential redundancy and noise introduced by dense all-to-all fusion between thousands of WSI patches and pathway tokens, as well as the limited ability of fixed graphs to capture patient-specific pathology–molecular associations.

In the main comparison, the proposed model achieved the highest average C-index among the evaluated baselines and a higher cohort-specific mean C-index than SurvPath in all five cohorts. The evidence was strongest for BLCA, where the paired fold-level comparison reached statistical significance. For the remaining cohorts, the mean differences were positive, but statistical power was limited by the five-fold evaluation protocol. The results therefore indicate average improvements with cohort-dependent effect sizes rather than uniformly significant gains, particularly for cohorts with small margins, such as BRCA and HNSC.

Because the selected patches were not annotated by pathologists, we do not assign specific cell types, tissue structures, or diagnostic interpretations to the highlighted regions. The visualization therefore demonstrates model transparency at the pathway–patch level, but not established clinical utility.

The ablation studies further clarified the source and limitations of the improvement. Dynamic Top-k selection with $k = 8$ provided the highest average performance, whereas random, static, dense, overly sparse, or overly connected graph structures were less effective, suggesting that both adaptivity and controlled sparsity are important for modeling pathology–omics interactions. The stage-wise ablation indicated that graph interaction was most useful when combined with sequence-level contextual modeling, whereas graph-only fusion was insufficient. The modality ablation further showed that omics-only and histology-only baselines could be competitive in individual cohorts, consistent with the biological heterogeneity of survival prediction across tumor types.

The efficiency analysis indicated that the proposed fusion module was practical under a matched profiling setting. Although our model contained more parameters than the Transformer replacement, it achieved lower peak GPU memory usage and comparable runtime under the same hardware, input size, and training configuration. These profiling results quantify computational cost and do not replace the full five-fold performance evaluation.

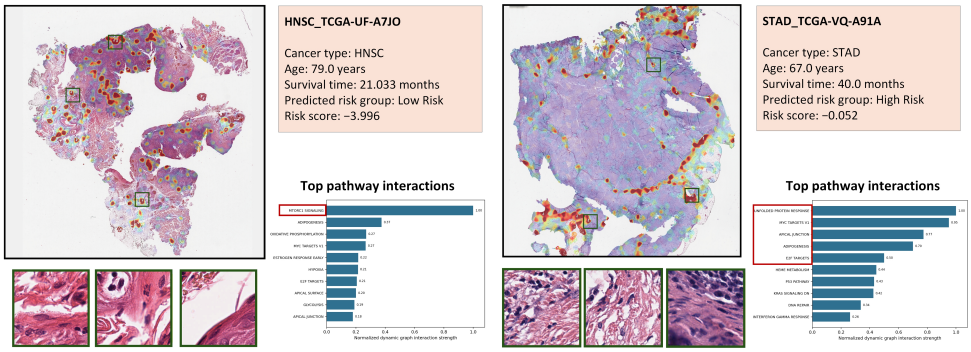


Fig. 4 Case-level pathway–patch interpretability visualization. WSI heatmaps indicate high-interaction tissue regions identified by the learned dynamic graph, and enlarged patches show local morphology. Bar plots rank Hallmark pathway tokens by normalized interaction strength. The HNSC case is associated with `MTORC1_SIGNALING`, while the STAD case is associated with `UNFOLDED_PROTEIN_RESPONSE`. These pathway–patch associations are model-derived hypotheses and require further pathological and biological validation.

Several limitations remain. First, all experiments are based on TCGA cohorts, and patient-level cross-validation does not establish robustness to institutional, scanner, staining, preprocessing, or molecular batch variations. Independent external validation on matched WSI-omics survival datasets and site-aware analyses are therefore needed. Second, the main results use random seed 1 and validation-selected hyperparameters from a predefined sweep; fold-level variability does not replace repeated-seed evaluation. Third, the pathway–patch visualizations have not undergone pathological or biological validation and should not be interpreted mechanistically. Finally, although the Transformer replacement and TransMIL-style efficiency controls broaden the empirical comparison, the main performance table does not include a dedicated recent graph-based WSI survival baseline. These issues limit claims of robustness, interpretability, and generalizability beyond the evaluated TCGA setting.

6 Conclusion

We presented a dynamic graph and multi-stage Mamba fusion framework for multimodal cancer survival prediction from WSIs and genomic pathways. The method constructs patient-specific sparse WSI–pathway graphs, refines selected cross-modal interactions using graph attention, and aggregates multimodal tokens through Mamba state-space blocks. Across five TCGA cohorts, the proposed model achieved the highest average C-index among the evaluated baselines, higher cohort-specific mean performance than SurvPath, and significant Kaplan–Meier risk stratification in four cohorts. The statistical and ablation analyses showed that the benefit was cohort-dependent, with the strongest support provided by the BLCA paired comparison and matched Top-k graph ablation. These results support sample-adaptive sparse multimodal interaction modeling as a practical strategy for WSI-omics survival prediction; however, external validation and pathologist-guided interpretation remain necessary before clinical deployment.

Declarations

Funding. This work was supported in part by the National Key Research and Development Program of China under Grant 2019YFB1311503.

Competing interests. The authors declare no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

References

[1] Harrell, F.E. Jr, Harrell, F.E.: Cox proportional hazards regression model. Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis, 475–519 (2015)

[2] Zhu, X., Yao, J., Huang, J.: Deep convolutional neural network for survival analysis with pathological images. In: 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 544–547 (2016). IEEE

[3] Chen, R.J., Lu, M.Y., Shaban, M., Chen, C., Chen, T.Y., Williamson, D.F., Mahmood, F.: Whole slide images are 2d point clouds: Context-aware survival prediction using patch-based graph convolutional networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 339–349 (2021)

[4] Zheng, Y., Gindra, R.H., Green, E.J., Burks, E.J., Betke, M., Beane, J.E., Kolachalama, V.B.: A graph-transformer for whole slide image classification. IEEE transactions on medical imaging **41**(11), 3003–3015 (2022)

[5] Lee, Y., Park, J.H., Oh, S., Shin, K., Sun, J., Jung, M., Lee, C., Kim, H., Chung, J.-H., Moon, K.C., et al.: Derivation of prognostic contextual histopathological features from whole-slide images of tumours via graph deep learning. Nature Biomedical Engineering, 1–15 (2022)

[6] Chan, T.H., Cendra, F.J., Ma, L., Yin, G., Yu, L.: Histopathology whole slide image analysis with heterogeneous graph representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15661–15670 (2023)

[7] Kumar, L., Greiner, R.: Gene expression based survival prediction for cancer patients—a topic modeling approach. PloS one **14**(11), 0224446 (2019)

[8] Hou, J., Aerts, J., Den Hamer, B., Van Ijcken, W., Den Bakker, M., Riegman, P., Leest, C., Spek, P., Foekens, J.A., Hoogsteden, H.C., *et al.*: Gene expression-based classification of non-small cell lung carcinomas and survival prediction. PloS one **5**(4), 10312 (2010)

[9] Huang, Z., Zhan, X., Xiang, S., Johnson, T.S., Helm, B., Yu, C.Y., Zhang, J., Salama, P., Rizkalla, M., Han, Z., *et al.*: Salmon: survival analysis learning with

multi-omics neural networks on breast cancer. Frontiers in genetics **10**, 166 (2019)

[10] Tas, F., Ciftci, R., Kilic, L., Karabulut, S.: Age is a prognostic factor affecting survival in lung cancer patients. Oncology letters **6**(5), 1507–1513 (2013)

[11] Katzman, J.L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., Kluger, Y.: Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. BMC medical research methodology **18**, 1–12 (2018)

[12] Qu, L., Ma, Y., Luo, X., Guo, Q., Wang, M., Song, Z.: Rethinking multiple instance learning for whole slide image classification: A good instance classifier is all you need. IEEE Transactions on Circuits and Systems for Video Technology (2024)

[13] Song, A.H., Chen, R.J., Jaume, G., Vaidya, A.J., Baras, A., Mahmood, F.: Multimodal prototyping for cancer survival prediction. In: Forty-first International Conference on Machine Learning

[14] 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)

[15] Zhou, F., Chen, H.: Cross-modal translation and alignment for survival analysis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 21485–21494 (2023)

[16] Li, C., Liu, X., Wang, C., Liu, Y., Yu, W., Shao, J., Yuan, Y.: Gtp-4o: Modality-prompted heterogeneous graph learning for omni-modal biomedical representation. In: European Conference on Computer Vision, pp. 168–187 (2025). Springer

[17] Wiegrebe, S., Kopper, P., Sonabend, R., Bischl, B., Bender, A.: Deep learning for survival analysis: a review. Artificial Intelligence Review **57**(3), 65 (2024)

[18] Le, V.H., Minh, T.N.T., Kha, Q.H., Le, N.Q.K.: Deep learning radiomics for survival prediction in non-small-cell lung cancer patients from ct images. Journal of Medical Systems **49**(1), 22 (2025) https://doi.org/10.1007/s10916-025-02156-5

[19] Le, V.H., Minh, T.N.T., Kha, Q.H., Le, N.Q.K.: A transfer learning approach on mri-based radiomics signature for overall survival prediction of low-grade and high-grade gliomas. Medical & Biological Engineering & Computing **61**(10), 2699–2712 (2023) https://doi.org/10.1007/s11517-023-02875-2

[20] Behrouz, A., Hashemi, F.: Graph mamba: Towards learning on graphs with state space models. In: Proceedings of the 30th ACM SIGKDD Conference on

Knowledge Discovery and Data Mining, pp. 119–130 (2024)

[21] Guo, H., Li, J., Dai, T., Ouyang, Z., Ren, X., Xia, S.-T.: Mambair: A simple baseline for image restoration with state-space model. In: European Conference on Computer Vision, pp. 222–241 (2025). Springer

[22] Rangapuram, S.S., Seeger, M.W., Gasthaus, J., Stella, L., Wang, Y., Januschowski, T.: Deep state space models for time series forecasting. Advances in neural information processing systems **31** (2018)

[23] Waleffe, R., Byeon, W., Riach, D., Norick, B., Korthikanti, V., Dao, T., Gu, A., Hatamizadeh, A., Singh, S., Narayanan, D., et al.: An empirical study of mamba-based language models. arXiv preprint arXiv:2406.07887 (2024)

[24] Ahamed, M.A., Cheng, Q.: Timemachine: A time series is worth 4 mambas for long-term forecasting. arXiv preprint arXiv:2403.09898 (2024)

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

[26] 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 (2024). Springer

[27] Fang, Z., Wang, Y., Zhang, Y., Wang, Z., Zhang, J., Ji, X., Zhang, Y.: Mammil: Multiple instance learning for whole slide images with state space models. In: 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 3200–3205 (2024). IEEE

[28] Fan, J., Lv, T., Di, Y., Li, L., Pan, X.: Pathmamba: Weakly supervised state space model for multi-class segmentation of pathology images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 500–509 (2024). Springer

[29] Gu, A., Goel, K., Gupta, A., Ré, C.: On the parameterization and initialization of diagonal state space models. Advances in Neural Information Processing Systems **35**, 35971–35983 (2022)

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

[31] 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)

[32] 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)

[33] Li, Z., Jiang, Y., Lu, M., Li, R., Xia, Y.: Survival prediction via hierarchical multimodal co-attention transformer: A computational histology-radiology solution. IEEE Transactions on Medical Imaging **42**(9), 2678–2689 (2023)

[34] Ding, K., Zhou, M., Metaxas, D.N., Zhang, S.: Pathology-and-genomics multimodal transformer for survival outcome prediction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 622–631 (2023). Springer

[35] Jaume, G., Vaidya, A., Chen, R.J., Williamson, D.F., Liang, P.P., Mahmood, F.: Modeling dense multimodal interactions between biological pathways and histology for survival prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11579–11590 (2024)

[36] Zheng, Y., Conrad, R.D., Green, E.J., Burks, E.J., Betke, M., Beane, J.E., Kolachalama, V.B.: Graph attention-based fusion of pathology images and gene expression for prediction of cancer survival. IEEE transactions on medical imaging (2024)

[37] Hou, W., Lin, C., Yu, L., Qin, J., Yu, R., Wang, L.: Hybrid graph convolutional network with online masked autoencoder for robust multimodal cancer survival prediction. IEEE Transactions on Medical Imaging **42**(8), 2462–2473 (2023)

[38] Liu, P., Ji, L., Ye, F., Fu, B.: Graphlsurv: A scalable survival prediction network with adaptive and sparse structure learning for histopathological whole-slide images. Computer Methods and Programs in Biomedicine **231**, 107433 (2023)

[39] Xu, B., Xu, C., Su, B.: Cross-modal graph attention network for entity alignment. In: Proceedings of the 31st ACM International Conference on Multimedia, pp. 3715–3723 (2023)

[40] Xiong, C., Chen, H., Zheng, H., Wei, D., Zheng, Y., Sung, J.J., King, I.: Mome: Mixture of multimodal experts for cancer survival prediction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 318–328 (2024). Springer

[41] 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)

[42] Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R., Ozenberger, B.A.,

Ellrott, K., Shmulevich, I., Sander, C., Stuart, J.M.: The cancer genome atlas pan-cancer analysis project. Nature genetics **45**(10), 1113–1120 (2013)

[43] 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)

[44] 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)

[45] 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)

[46] Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International Conference on Machine Learning, pp. 2127–2136 (2018). PMLR

[47] Yao, J., Zhu, X., Jonnagaddala, J., Hawkins, N., Huang, J.: Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. Medical Image Analysis **65**, 101789 (2020)

[48] 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)