MSSF-SURV: MULTIMODAL STATE-SPACE FUSION FOR CANCER SURVIVAL PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Accurate cancer survival prediction informs patient risk stratification and individualized treatment planning. Histopathology and multi-omics provide complementary prognostic evidence, but their integration can be limited by the loss of tissue context during patch aggregation and insufficient modeling of cross-modal interactions. We propose Multimodal State Space Fusion (MSSF-Surv), a framework that preserves spatial organization before conditioning pathology readout on molecular profiles. Its dual-granularity design captures tissue architecture alongside cellular morphology, while multidirectional state-space modeling handles long sequences of fine-grained whole-slide image patches with linear complexity. Molecular queries then retrieve information from these spatially contextualized representations, rather than from isolated patch features. A shared state-space model combines the retrieved features with omics representations through modality-specific gates, linking tissue context to molecular information for survival prediction. Extensive experiments using five-fold cross-validation on six TCGA cancer cohorts demonstrate the effectiveness of MSSF-Surv. Survival analyses and ablations support its predictive performance and the contributions of dual-granularity encoding and omics-guided fusion.

1 INTRODUCTION

Accurate cancer survival prediction supports risk stratification and individualized treatment planning. Histopathology and multi-omics offer complementary views of tumor biology: whole-slide images (WSIs) reveal cellular morphology and tissue architecture, while molecular profiles characterize underlying biological variation (Song et al., 2023; Schneider et al., 2022). Their integration offers an opportunity to relate molecular information to its tissue context. Realizing this potential requires representations that preserve how local morphology is organized across a slide and allow molecular information to guide the interpretation of these spatial patterns.

This requirement is difficult to meet at gigapixel resolution. A WSI can contain tens to hundreds of thousands of patches, commonly encoded independently and aggregated through multiple-instance learning (Chen et al., 2021; 2022a; Wang et al., 2023; Liu et al., 2024; Yao et al., 2020). Pooling without spatial information cannot explicitly distinguish different arrangements of similar local features, although tumor–stroma organization and regional heterogeneity may carry prognostic information. Self-attention can model dependencies across patches, but its quadratic complexity makes long sequences expensive (Li et al., 2024). Reducing image resolution eases this burden at the cost of fine morphological detail. The challenge is therefore to retain both tissue architecture and cellular detail while modeling context across distant regions.

Multimodal learning adds a related challenge: molecular guidance can incorporate tissue context when the pathology features being selected encode their spatial surroundings. Existing approaches have advanced beyond feature concatenation through omics-guided attention and cross-modal alignment (Chen et al., 2021; Zhou & Chen, 2023; Wang et al., 2025). Yet attention over independently encoded patches primarily selects local appearances; it does not itself establish the spatial relationships needed to interpret them. Intramodal context modeling and cross-modal interaction must therefore be considered together, with attention to both pathological scale and computational cost.

We propose Multimodal State-Space Fusion (MSSF-Surv) around this principle. The framework builds spatial context at two granularities before applying molecular guidance. A Pathology-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Transformer encodes coarse tissue architecture, while a multidirectional Pathology-Mamba propagates context along fine-grained patch sequences. State-space updates make the latter linear in sequence length (Gu et al., 2021; Liu et al., 2025), allowing contextual encoding without replacing fine patches with a coarser representation. Omics queries then retrieve features from both contextualized pathology streams. The resulting readout can reflect regional organization as well as individual patch appearance. A shared state-space model integrates the retrieved features with omics representations, and stream-specific gates modulate its output according to the input streams. This design connects molecularly guided selection to tissue organization across scales.

Our main contributions are as follows:

- We introduce dual-granularity spatial encoding for multimodal survival prediction. By pairing coarse architectural context with efficient long-range modeling of fine morphology, MSSF-Surv supplies molecular queries with pathology representations that retain both scales of tissue organization.
- We develop context-aware, omics-guided fusion that combines molecularly selected pathology features and omics representations through shared state-space dynamics and stream-specific gates. This mechanism couples cross-modal selection with adaptive integration of the three streams.
- We evaluate MSSF-Surv using five-fold cross-validation on six public TCGA cancer cohorts. It achieves the highest mean C-index on five cohorts and comparable performance on LGG. Survival analyses and ablations further support its predictive performance and the contributions of dual-granularity encoding and omics-guided fusion.

2 Related Work

2.1 Long-range Dependency Modeling for Whole Slide Images

Multiple-instance learning (MIL) represents a WSI as a bag of patches, enabling learning from slide-level supervision. Attention-based pooling (Ilse et al., 2018; Lu et al., 2021) weights informative patches without explicitly encoding their spatial arrangement. Graph models (Wang et al., 2024c) and TransMIL (Shao et al., 2021) model relationships among patches, while HIPT and HC-MAE capture tissue information across scales through hierarchical representations (Chen et al., 2022a; Wang et al., 2023).

Mamba provides selective state-space modeling with linear sequence complexity (Gu & Dao, 2023). PathMamba applies this approach to pathology segmentation (Fan et al., 2024), and MambaMIL models long patch sequences through sequence reordering (Yang et al., 2024).

Preserving tissue architecture and cellular detail requires spatial context at both scales, but attention over long fine-grained sequences is costly. MSSF-Surv combines dual-granularity spatial encoding with linear-complexity fine-grained modeling, preserving both scales before omics-guided selection.

2.2 Multimodal Pathology–Omics Fusion for Cancer Survival Prediction

Pathology–omics methods differ in where modalities interact. PORPOISE combines separately encoded histology and molecular profiles (Chen et al., 2022b), whereas Pathomic Fusion uses tensor products to represent feature interactions (Chen et al., 2020). MCAT uses genomic queries to aggregate pathology features (Chen et al., 2021). Omics-guided cross-attention also appears in cloud-edge survival modeling (Wang et al., 2024b).

CMTA learns cross-modal translation and alignment through parallel encoder–decoder branches (Zhou & Chen, 2023). PathOmics aligns pathology and genomics representations through multimodal pretraining (Ding et al., 2023), while ContraMAE combines contrastive alignment with reconstruction objectives (Wang et al., 2024a). AlignMamba and FusionMamba extend state-space modeling to cross-modal alignment and feature fusion (Li et al., 2025; Xie et al., 2024).

Cross-modal alignment alone does not determine how tissue information at different scales should contribute to prediction. MSSF-Surv links omics-guided selection from spatially contextualized fea-

tures to shared state-space fusion, using stream-specific gates to adaptively integrate coarse pathology, fine pathology, and molecular evidence.

3 Multimodal State-Space Fusion Framework

MSSF-Surv predicts patient survival risk by combining spatially contextualized pathology features with molecular information (Fig. 1). Let $\mathcal{X} = \{(\mathbf{P}_i, \mathbf{O}_i, t_i, e_i)\}_{i=1}^{L}$ denote the training cohort, where $\mathbf{P}_i$ is a WSI, $\mathbf{O}_i$ contains multi-omics measurements, and t_i is the observed time to an event or censoring. The event indicator is $e_i = 1$ for an observed event and $e_i = 0$ for a censored observation. Each slide provides N fine patches of 256×256 pixels and M coarse patches of 4096×4096 pixels. The preprocessed omics input contains three modalities with 300 features each.

Pathology and omics encoders provide features for molecularly guided readout and shared state-space fusion. A linear head maps the pooled representation to risk. We omit patient and batch indices unless needed and represent tokens as rows.

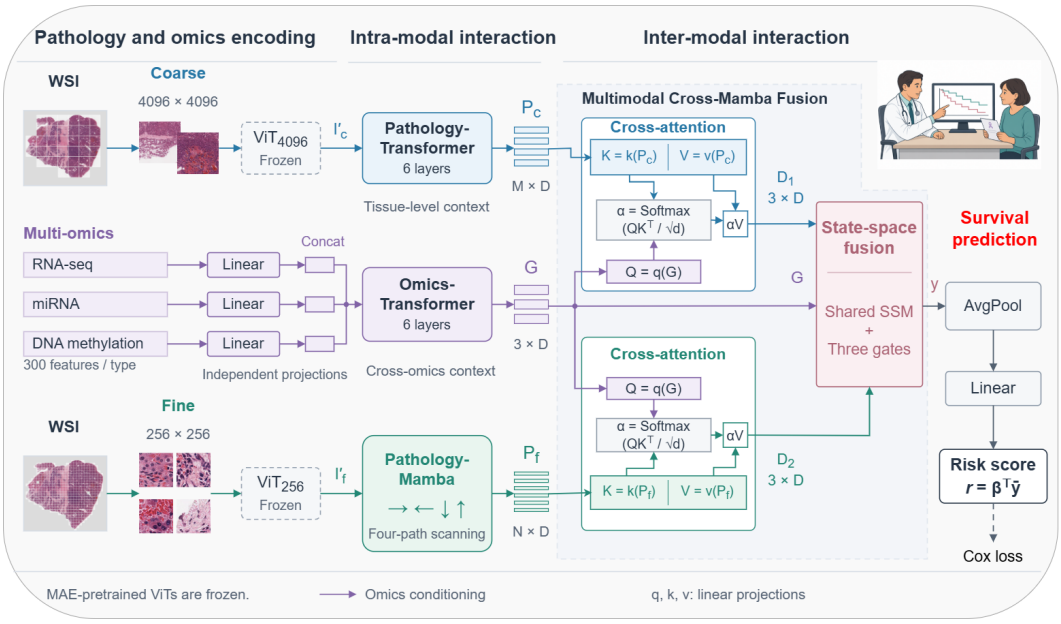


Figure 1: Overview of MSSF-Surv. MSSF-Surv couples dual-granularity spatial context with omics-guided fusion, linking tissue architecture and cellular morphology to molecular profiles through shared state-space dynamics.

3.1 Dual-Granularity Representation Learning for Pathology Images

The pathology branches retain context at two scales: self-attention models relationships among coarse tissue regions, while multidirectional state-space processing captures dependencies along the much longer fine-grained patch sequence.

Separate Vision Transformers, pretrained by masked autoencoding and then frozen, produce embeddings at the two patch resolutions:

$$\mathbf{I}'_c = \mathrm{ViT}_{4096}(\mathbf{I}_c), \qquad \mathbf{I}'_c \in \mathbb{R}^{M \times D}. \tag{1}$$

$$\mathbf{I}'_f = \mathrm{ViT}_{256}(\mathbf{I}_f), \qquad \mathbf{I}'_f \in \mathbb{R}^{N \times D}. \tag{2}$$

Here, D is the shared embedding dimension, and each row represents one patch.

The Pathology-Transformer processes coarse embeddings to model dependencies among tissue regions:

$$\mathbf{P}_c = \mathrm{Transformer}_{\mathrm{path}}(\mathbf{I}'_c), \qquad \mathbf{P}_c \in \mathbb{R}^{M \times D}. \tag{3}$$

Algorithm 1 Four-path scanning Pathology-Mamba block

Input: Fine-grained pathology patch feature sequence $\mathbf{I}'_f : (B, N, D)$, and four scanning path permutations s_1, s_2, s_3, s_4 of length N.
Output: Fine-grained pathology representation $\mathbf{P}_f : (B, N, D)$

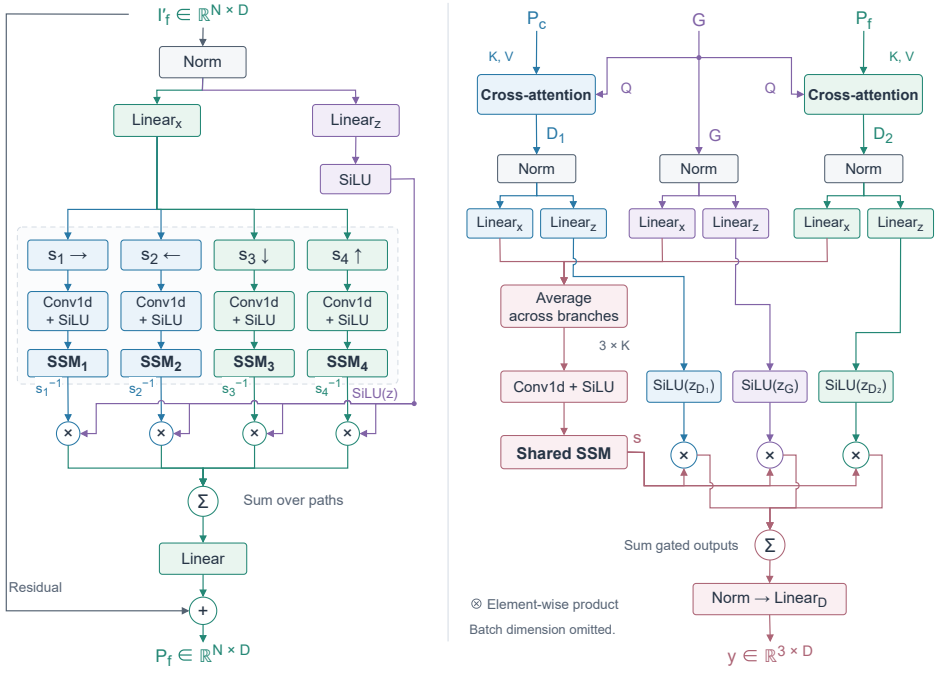
$\mathbf{I}_f^{\text{Norm}} : (B, N, D) \leftarrow \text{Normalize}(\mathbf{I}'_f)$
$\mathbf{x} : (B, N, K) \leftarrow \text{Linear}_x(\mathbf{I}_f^{\text{Norm}})$
$\mathbf{z} : (B, N, K) \leftarrow \text{Linear}_z(\mathbf{I}_f^{\text{Norm}})$
for $j \in \{1, 2, 3, 4\}$ **do**
  $\mathbf{x}'_j : (B, N, K) \leftarrow \text{SiLU}(\text{Conv1d}_j(\mathbf{x}[:, s_j, :]))$
  $\mathbf{A}_j : (K, E) \leftarrow \text{Parameter}_j^A$ {Initialized with HiPPO}
  $\mathbf{B}_j : (B, N, E) \leftarrow \text{Linear}_j^B(\mathbf{x}'_j)$
  $\mathbf{C}_j : (B, N, E) \leftarrow \text{Linear}_j^C(\mathbf{x}'_j)$
  $\mathbf{\Delta}_j : (B, N, K) \leftarrow \text{Softplus}(\text{Linear}_j^{\Delta}(\mathbf{x}'_j) + \text{Parameter}_j^{\Delta})$
  $\overline{\mathbf{A}}_j, \overline{\mathbf{B}}_j : (B, N, K, E) \leftarrow \text{Discretize}(\mathbf{\Delta}_j, \mathbf{A}_j, \mathbf{B}_j)$
  $\mathbf{y}_j : (B, N, K) \leftarrow \text{SSM}(\overline{\mathbf{A}}_j, \overline{\mathbf{B}}_j, \mathbf{C}_j)(\mathbf{x}'_j)$
  $\mathbf{y}'_j : (B, N, K) \leftarrow \mathbf{y}_j[:, s_j^{-1}, :] \odot \text{SiLU}(\mathbf{z})$ {Inverse-permutation gating}
end for
$\mathbf{P}_f : (B, N, D) \leftarrow \text{Linear}^I(\sum_{j=1}^{4} \mathbf{y}'_j) + \mathbf{I}'_f$

Each coarse patch retains a contextualized token for molecularly guided aggregation.

Pathology-Mamba processes the fine embeddings along four spatial scanning paths (Figs. 2a and 3):

$$\mathbf{P}_f = \text{Mamba}_{\text{path}}(\mathbf{I}'_f), \qquad \mathbf{P}_f \in \mathbb{R}^{N \times D}. \tag{4}$$

Algorithm 1 details one Pathology-Mamba block. In both algorithms, B is the batch size, K is the expanded feature width, and E is the state dimension. Each path permutes the patch sequence; its output is restored to the original order before gating and summation, so corresponding spatial locations are combined.



(a) Four-path Pathology-Mamba (b) Multimodal Cross-Mamba Fusion

Figure 2: Pathology-Mamba block (a) and Multimodal Cross-Mamba Fusion (b). Inverse permutations restore patch order before gating; stream-specific gates modulate the shared SSM output.

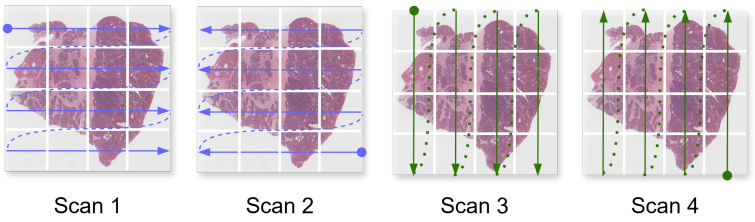


Figure 3: Four directional scanning paths used in Pathology-Mamba.

3.2 Multi-omics Data Representation Learning

The omics input comprises RNA-seq expression, miRNA expression, and DNA methylation. For each modality m, preprocessing yields a vector $\mathbf{O}^m_{\text{pre}} \in \mathbb{R}^{1\times 300}$. Separate linear projections accommodate the different feature sets while mapping them to the common embedding dimension:

$$\mathbf{H}_m = \text{Linear}_m(\mathbf{O}^m_{\text{pre}}), \qquad m \in \{\text{rna}, \text{mirna}, \text{dna_meth}\}. \tag{5}$$

Each embedding $\mathbf{H}_m \in \mathbb{R}^{1\times D}$ represents an omics modality rather than an individual molecular feature.

We stack these embeddings along the token dimension and apply an Omics-Transformer:

$$\mathbf{G} = \text{Transformer}_{\text{omics}}([\mathbf{H}_{\text{rna}}; \mathbf{H}_{\text{mirna}}; \mathbf{H}_{\text{dna_meth}}]), \qquad \mathbf{G} \in \mathbb{R}^{3\times D}. \tag{6}$$

The three contextualized tokens retain a fixed modality order and query both pathology branches. They capture associations between tissue and omics modalities without assuming one-to-one correspondence between individual genes and patches.

3.3 Multimodal Cross-Mamba Fusion

Fusion combines molecularly guided readout with shared state-space integration (Fig. 2b). Cross-attention reduces each pathology sequence to three tokens aligned with the omics queries, so fusion operates on corresponding query positions.

For coarse pathology, omics tokens provide queries and contextualized tissue tokens provide keys and values:

$$\mathbf{D}_1 = \text{Softmax}\left(\frac{(\mathbf{G}\mathbf{W}_q)(\mathbf{P}_c\mathbf{W}_k)^\top}{\sqrt{d}}\right)(\mathbf{P}_c\mathbf{W}_v). \tag{7}$$

Here, $\mathbf{W}_q, \mathbf{W}_k \in \mathbb{R}^{D\times d}$ and $\mathbf{W}_v \in \mathbb{R}^{D\times D}$ project features. d denotes query/key width. Softmax acts over pathology tokens. The attention matrix has shape $3 \times M$, yielding $\mathbf{D}_1 \in \mathbb{R}^{3\times D}$. Its rows summarize coarse tissue context conditioned on the corresponding molecular queries.

The fine branch uses the same query representation to retrieve cellular morphological information:

$$\mathbf{D}_2 = \text{Softmax}\left(\frac{(\mathbf{G}\mathbf{W}_q)(\mathbf{P}_f\mathbf{W}_k)^\top}{\sqrt{d}}\right)(\mathbf{P}_f\mathbf{W}_v). \tag{8}$$

The attention matrix now has shape $3 \times N$, while $\mathbf{D}_2 \in \mathbb{R}^{3\times D}$. Corresponding rows of the two readouts refer to the same omics query, although they aggregate pathology at different granularities.

We normalize $\mathbf{D}_1$, $\mathbf{D}_2$, and $\mathbf{G}$ and project each stream into state inputs and gates of width K. Averaging occurs across the three streams at corresponding token positions, preserving the three-token sequence. A one-dimensional convolution followed by SiLU produces the shared input $\mathbf{x}$. Input-dependent projections generate the selection parameters $\mathbf{B}$, $\mathbf{C}$, and positive step sizes $\mathbf{\Delta}$.

Discretizing $\mathbf{A}$ and $\mathbf{B}$ with $\mathbf{\Delta}$ yields the selective state-space operator. Its output $\mathbf{s} \in \mathbb{R}^{3\times K}$ represents sequence features, not the internal recurrent state. Each stream's SiLU-activated gate modulates this shared output. We sum the gated outputs, normalize, and project to $\mathbf{y} \in \mathbb{R}^{3\times D}$ (Algorithm 2). Thus, averaged inputs condition shared state-space dynamics, while separate gates modulate the resulting sequence without normalized attention weights.

Algorithm 2 Multimodal Cross-Mamba Fusion (MCMF)

Input: Coarse- and fine-grained pathology representations $\mathbf{P}_c : (B, M, D)$ and $\mathbf{P}_f : (B, N, D)$, and multi-omics data representation $\mathbf{G} : (B, 3, D)$.
Output: Multimodal fusion representation $\mathbf{y} : (B, 3, D)$

$\mathbf{D}_1 : (B, 3, D) \leftarrow \mathrm{CrossAttn}(\mathbf{G}\mathbf{W}_q, \mathbf{P}_c\mathbf{W}_k, \mathbf{P}_c\mathbf{W}_v)$
$\mathbf{D}_2 : (B, 3, D) \leftarrow \mathrm{CrossAttn}(\mathbf{G}\mathbf{W}_q, \mathbf{P}_f\mathbf{W}_k, \mathbf{P}_f\mathbf{W}_v)$
$\mathbf{D}'_1, \mathbf{D}'_2, \mathbf{G}' : (B, 3, D) \leftarrow \mathrm{Normalize}(\mathbf{D}_1), \mathrm{Normalize}(\mathbf{D}_2), \mathrm{Normalize}(\mathbf{G})$
$\mathbf{x}_{D_1}, \mathbf{z}_{D_1} : (B, 3, K) \leftarrow \mathrm{Linear}_{x,D_1}(\mathbf{D}'_1), \mathrm{Linear}_{z,D_1}(\mathbf{D}'_1)$
$\mathbf{x}_{D_2}, \mathbf{z}_{D_2} : (B, 3, K) \leftarrow \mathrm{Linear}_{x,D_2}(\mathbf{D}'_2), \mathrm{Linear}_{z,D_2}(\mathbf{D}'_2)$
$\mathbf{x}_G, \mathbf{z}_G : (B, 3, K) \leftarrow \mathrm{Linear}_{x,G}(\mathbf{G}'), \mathrm{Linear}_{z,G}(\mathbf{G}')$
$\mathbf{x} : (B, 3, K) \leftarrow \mathrm{SiLU}(\mathrm{Conv1d}((\mathbf{x}_{D_1} + \mathbf{x}_{D_2} + \mathbf{x}_G)/3))$
$\mathbf{A} : (K, E) \leftarrow \mathrm{Parameter}^{\mathrm{A}}$ {Initialized with HiPPO}
$\mathbf{B} : (B, 3, E) \leftarrow \mathrm{Linear}_{\mathrm{B}}(\mathbf{x})$
$\mathbf{C} : (B, 3, E) \leftarrow \mathrm{Linear}_{C}(\mathbf{x})$
$\mathbf{\Delta} : (B, 3, K) \leftarrow \mathrm{Softplus}(\mathrm{Linear}_{\Delta}(\mathbf{x}) + \mathrm{Parameter}_{\Delta})$
$\overline{\mathbf{A}}, \overline{\mathbf{B}} : (B, 3, K, E) \leftarrow \mathrm{Discretize}(\mathbf{\Delta}, \mathbf{A}, \mathbf{B})$
$\mathbf{s} : (B, 3, K) \leftarrow \mathrm{SSM}(\overline{\mathbf{A}}, \overline{\mathbf{B}}, \mathbf{C})(\mathbf{x})$
$\mathbf{y}' : (B, 3, K) \leftarrow \mathbf{s} \odot \mathrm{SiLU}(\mathbf{z}_{D_1}) + \mathbf{s} \odot \mathrm{SiLU}(\mathbf{z}_{D_2}) + \mathbf{s} \odot \mathrm{SiLU}(\mathbf{z}_G)$
$\mathbf{y} : (B, 3, D) \leftarrow \mathrm{Linear}_{D}(\mathrm{Normalize}(\mathbf{y}'))$

3.4 Survival Prediction

Average pooling over the three tokens gives the patient representation $\overline{\mathbf{y}}^{(i)} \in \mathbb{R}^D$. A linear head produces the log-risk score $r_i = \boldsymbol{\beta}^\top \overline{\mathbf{y}}^{(i)}$. Training minimizes the event-normalized negative Cox partial log-likelihood:

$$\mathcal{L}_{\mathrm{surv}} = -\frac{1}{n_E} \sum_{i:e_i=1} \left[r_i - \log \sum_{j:t_j \geq t_i} \exp(r_j) \right]. \tag{9}$$

Here, $n_E = \sum_i e_i$ counts observed events, and the denominator includes patients still at risk at time t_i. Censored patients contribute to eligible risk sets but do not contribute event terms. Larger scores indicate higher relative risk. The head does not directly estimate survival probabilities.

4 Experiments

We implement MSSF-Surv in PyTorch and run experiments on three NVIDIA A100 GPUs with 80 GB memory each. MAE-pretrained ViT encoders remain frozen during survival training and extract features at the two patch resolutions. Pathology-Mamba contains two Mamba blocks for fine-grained sequence modeling. Pathology-Transformer and Omics-Transformer each contain six layers, with a hidden dimension of 384 and 12 attention heads. The fusion module uses one Mamba block. We train for 80 epochs with Adam, using a learning rate of 5×10^{-4} and weight decay of 10^{-2}. Gradient accumulation spans 30 samples to accommodate variable-length WSI inputs under memory constraints.

4.1 Datasets and Evaluation Metrics

Datasets. We evaluate on six TCGA cohorts: breast invasive carcinoma (BRCA), lung adenocarcinoma (LUAD), liver hepatocellular carcinoma (LIHC), colon adenocarcinoma (COAD), stomach adenocarcinoma (STAD), and lower-grade glioma (LGG). Each cohort provides paired WSIs and molecular profiles comprising gene expression, miRNA expression, and DNA methylation.

WSI preprocessing uses CLAM to identify tissue regions. We partition the selected regions into coarse patches of 4096×4096 pixels and fine patches of 256×256 pixels. Four spatial scanning orders are constructed for each fine-grained sequence following Fig. 3.

Omics preprocessing handles missing values separately for each data type. For RNA-seq and miRNA expression, we remove features with missing entries. For DNA methylation, we impute missing values using K-nearest neighbors. Zero-variance features are removed from all modalities.

The preprocessing pipeline further includes differential expression analysis with PyDESeq2 and feature filtering at $|\log_2 FoldChange| > 1$ and $p < 0.05$. A Random Survival Forest ranks features by their association with survival, and the top 300 features from each modality form the omics input.

Evaluation Metrics. We use five-fold cross-validation within each cohort and measure discrimination with the concordance index (C-index). It compares predicted risk ordering with observed survival ordering:

$$\text{C-index} = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \mathbb{I}\left(\boldsymbol{\beta}^\top \overline{\mathbf{y}}^{(i)} > \boldsymbol{\beta}^\top \overline{\mathbf{y}}^{(j)}\right) \tag{10}$$

Here, $\mathbb{I}(\cdot)$ denotes the indicator function, and $\mathcal{P} = \{(i,j) \mid t_i < t_j, e_i = 1\}$ contains comparable patient pairs. The earlier observed time must correspond to an event. Higher C-index values indicate better risk ranking. Tables report the mean $\pm$ standard deviation across the five folds; error bars in the ablation figures show the same standard deviation.

4.2 Comparison with State-of-the-Art Methods

Table 1 compares MSSF-Surv with GPDBN, CAMR, DeepCorrSurv, MCAT, CMTA, and PathOmics. MSSF-Surv achieves the highest mean C-index on five of six cohorts. Its largest margin over the best comparator is 4.8 percentage points on LIHC (0.759 versus 0.711 for MCAT). On LGG, PathOmics is slightly higher (0.848 versus 0.847). The ablations below examine the contributions of individual design choices.

Table 1: Performance comparison of MSSF-Surv and state-of-the-art methods on six cancer datasets (mean $\pm$ standard deviation over five folds; higher is better).

Methods	**COAD**	**LIHC**	**STAD**	**BRCA**	**LUAD**	**LGG**
GPDBN	0.593 ± 0.081	0.643 ± 0.019	0.587 ± 0.025	0.636 ± 0.047	0.615 ± 0.053	0.844 ± 0.025
CAMR	0.606 ± 0.074	0.691 ± 0.052	0.587 ± 0.029	0.656 ± 0.072	0.647 ± 0.059	0.803 ± 0.044
DeepCorrSurv	0.549 ± 0.026	0.700 ± 0.048	0.609 ± 0.049	0.659 ± 0.018	0.662 ± 0.032	0.828 ± 0.034
MCAT	0.691 ± 0.132	0.711 ± 0.029	0.622 ± 0.034	0.663 ± 0.041	0.664 ± 0.026	0.844 ± 0.032
CMTA	0.641 ± 0.055	0.708 ± 0.034	0.631 ± 0.039	0.680 ± 0.063	0.670 ± 0.046	0.840 ± 0.038
PathOmics	0.629 ± 0.042	0.690 ± 0.013	0.622 ± 0.034	0.694 ± 0.074	0.662 ± 0.027	$\mathbf{0.848 \pm 0.032}$
MSSF-Surv (Ours)	$\mathbf{0.721 \pm 0.068}$	$\mathbf{0.759 \pm 0.034}$	$\mathbf{0.648 \pm 0.052}$	$\mathbf{0.711 \pm 0.048}$	$\mathbf{0.692 \pm 0.062}$	0.847 ± 0.056

4.3 Survival Analysis

We assess risk stratification by dividing patients at the cohort-specific median predicted risk score. Kaplan-Meier curves summarize survival in the resulting high- and low-risk groups, and log-rank tests assess differences between them. Figure 4 reports p-values below 0.05 for all six cohorts. Separation is particularly pronounced in LIHC and LGG, with p-values of 1.18×10^{-8} and 3.52×10^{-8}, respectively. The STAD curves approach one another late in follow-up, so separation is not uniform over time. These analyses complement the C-index by showing group-level survival differences associated with the predicted risk scores.

4.4 Ablation Study

We examine input composition and interaction modeling separately. Modality ablations compare single, paired, and complete inputs, while module ablations compare encoding alone with intra-modal interaction, inter-modal interaction, and their combination.

Modality ablation. Table 2 reports results for fine-grained pathology, coarse-grained pathology, and omics, individually and in combination. Paired and complete configurations use the same shared-SSM fusion module. Fine-grained pathology is the strongest individual input on all six cohorts. Coarse pathology exceeds omics on five cohorts, with STAD as the exception. The complete model exceeds the best single or paired configuration by 1.4, 1.7, 0.1, 0.7, 0.4, and 0.5 percentage points on COAD, LIHC, STAD, BRCA, LUAD, and LGG, respectively. Gains are largest on COAD and

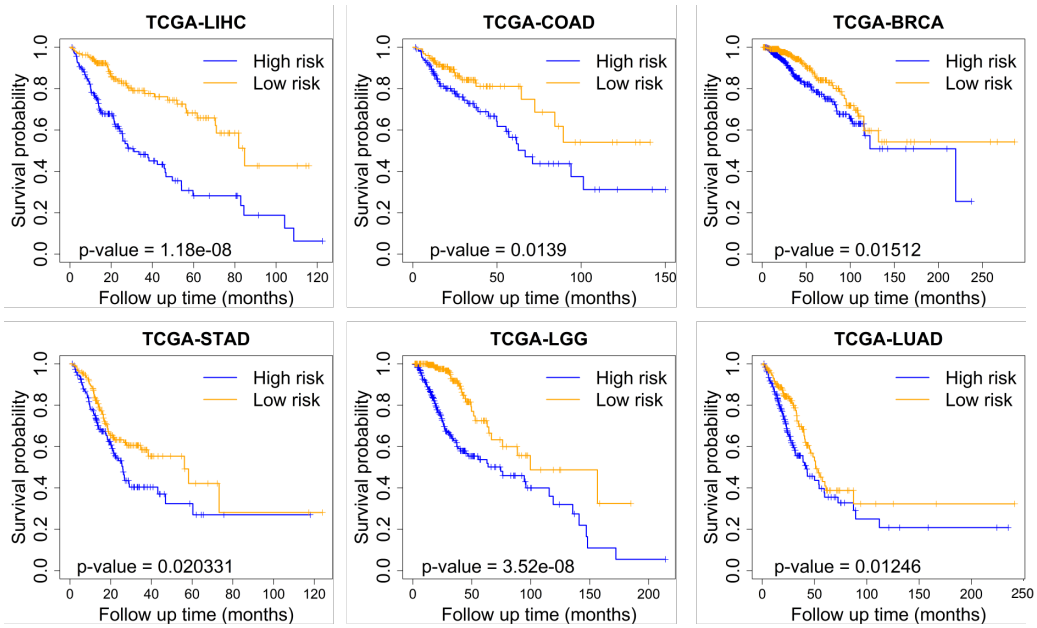


Figure 4: Survival analysis using Kaplan-Meier curves.

LIHC and smallest on STAD. These results support the complementary prognostic value of the three inputs, with their joint integration yielding the highest mean C-index on all six cohorts.

Among paired inputs, combining the two pathology scales gives the highest mean on five cohorts. Fine-grained pathology with omics performs best on STAD. The complete configuration achieves a higher mean C-index than all pairings across six cohorts, supporting the benefit of combining tissue architecture, cellular morphology, and molecular profiles.

Table 2: Ablation experiment of modality (C-index, mean ± standard deviation over five folds).

Input streams			C-index					
Fine-grained	Coarse-grained	Omics	**COAD**	**LIHC**	**STAD**	**BRCA**	**LUAD**	**LGG**
✓	–	–	0.696 ± 0.046	0.732 ± 0.038	0.640 ± 0.053	0.703 ± 0.053	0.680 ± 0.045	0.838 ± 0.053
–	✓	–	0.671 ± 0.052	0.718 ± 0.053	0.635 ± 0.026	0.701 ± 0.022	0.675 ± 0.072	0.833 ± 0.036
–	–	✓	0.655 ± 0.063	0.705 ± 0.051	0.636 ± 0.043	0.692 ± 0.043	0.653 ± 0.038	0.823 ± 0.066
✓	✓	–	0.707 ± 0.038	0.742 ± 0.043	0.643 ± 0.063	0.704 ± 0.047	0.688 ± 0.045	0.842 ± 0.073
✓	–	✓	0.699 ± 0.061	0.739 ± 0.037	0.647 ± 0.035	0.702 ± 0.058	0.683 ± 0.063	0.836 ± 0.046
–	✓	✓	0.686 ± 0.055	0.736 ± 0.072	0.639 ± 0.055	0.700 ± 0.063	0.680 ± 0.025	0.834 ± 0.051
✓	✓	✓	**0.721 ± 0.068**	**0.759 ± 0.034**	**0.648 ± 0.052**	**0.711 ± 0.048**	**0.692 ± 0.062**	**0.847 ± 0.056**

Module ablation. Figure 5 compares four configurations built on feature encoding. Adding either intra-modal or inter-modal interaction improves the mean C-index over encoding alone on every cohort. Inter-modal interaction exceeds intra-modal interaction on five cohorts, whereas LUAD slightly favors intra-modal interaction (0.677 versus 0.676). Combining both gives the highest mean throughout, exceeding the stronger individual interaction variant by 0.9–2.4 percentage points. The pattern supports modeling tissue and omics context together with cross-modal exchange. The relative benefit of each interaction type nevertheless varies by cohort.

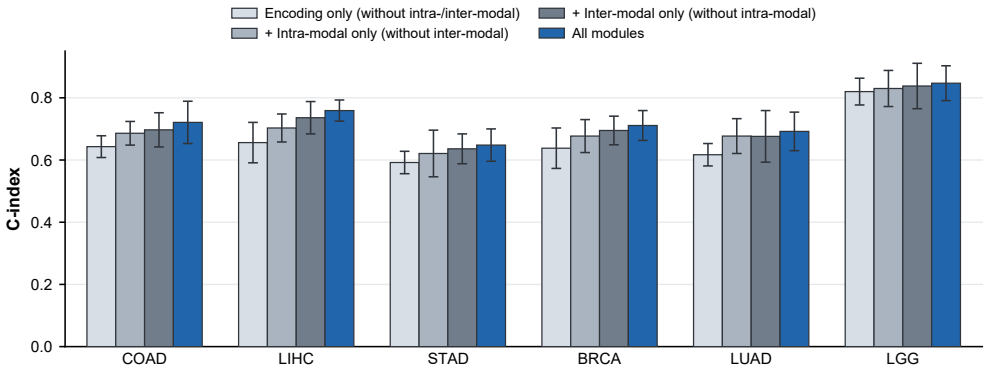


Figure 5: Module ablation on six TCGA cohorts. Bars and error bars show the mean C-index and standard deviation over five folds. Combining intra-modal and inter-modal interaction gives the highest mean on each cohort.

4.5 Impact of Cross-Attention on Survival Prediction Performance

We compare multi-omics, coarse-grained pathology, and fine-grained pathology queries (Fig. 6). Omics queries yield the highest mean C-index on five cohorts, improving over fine-grained queries by 0.1–0.6 percentage points. STAD favors fine-grained queries (0.652 versus 0.648). Coarse-grained queries perform below both alternatives on every cohort. These results support omics-guided readout as the default, with cohort-dependent benefits. Error bars overlap across query configurations.

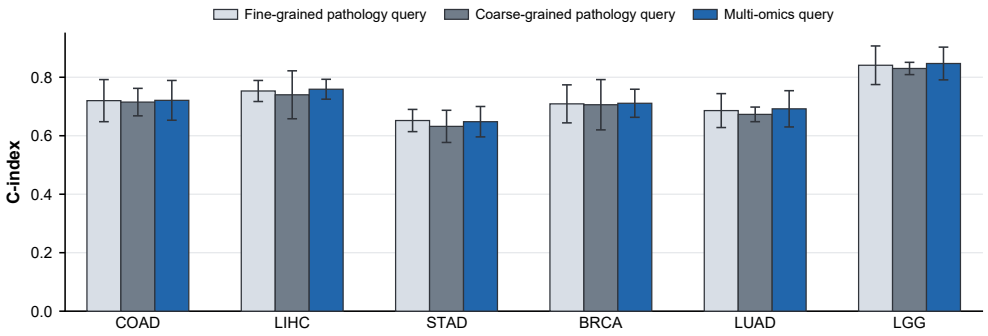


Figure 6: Cross-attention query ablation on six TCGA cohorts. Bars and error bars show the mean C-index and standard deviation over five folds. Omics queries give the highest mean on five cohorts; fine-grained pathology queries perform best on STAD.

5 Conclusion

We introduced MSSF-Surv for cancer survival prediction by connecting spatially contextualized pathology with molecular information. Dual-granularity encoding preserves tissue architecture and cellular morphology, with multidirectional state-space processing capturing long-range context in fine-grained patch sequences at linear cost in sequence length. Omics-guided readout then selects from these contextualized features, and shared state-space fusion with modality-specific gates integrates the three streams. On six TCGA cohorts, MSSF-Surv achieves the highest mean C-index on five and performs comparably to PathOmics on LGG. Ablations support the complementary value of the input modalities and of intra-modal and inter-modal interactions. Together, these findings support linking tissue context across scales to molecularly guided integration for survival prediction.

AI USE STATEMENT

In this work, we used generative AI tools solely to polish author-written text, including correcting grammar and improving clarity and readability. We did not use generative AI tools for literature retrieval or discovery, research ideation or execution, drafting sections of the paper, generating synthetic datasets, or formulating or proving mathematical claims. We have reviewed all AI-assisted edits to ensure that they preserve the intended meaning and accurately reflect our methods, results, and conclusions. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We have made efforts to ensure the reproducibility of this work. The architecture and key components of MSSF-Surv are described in the methodology section, while the datasets, preprocessing procedures, cross-validation protocol, evaluation metrics, and training configurations are detailed in the experimental section. The source code will be released upon acceptance.

REFERENCES

Richard J Chen, Ming Y Lu, Jingwen Wang, Drew FK Williamson, Scott J Rodig, Neal I Lindeman, and Faisal Mahmood. Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Transactions on Medical Imaging*, 41 (4):757–770, 2020.

Richard J Chen, Ming Y Lu, Wei-Hung Weng, Tiffany Y Chen, Drew FK Williamson, Trevor Manz, Maha Shady, and Faisal Mahmood. 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.

Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, and Faisal 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*, pp. 16144–16155, 2022a.

Richard J Chen, Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Jana Lipkova, Zahra Noor, Muhammad Shaban, Maha Shady, Mane Williams, Bumjin Joo, et al. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell*, 40(8):865–878, 2022b.

Kexin Ding, Mu Zhou, Dimitris N Metaxas, and Shaoting Zhang. Pathology-and-genomics multimodal transformer for survival outcome prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 622–631. Springer, 2023.

Jiansong Fan, Tianxu Lv, Yicheng Di, Lihua Li, and Xiang Pan. 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. Springer, 2024.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International Conference on Machine Learning*, pp. 2127–2136. PMLR, 2018.

Honglin Li, Yunlong Zhang, Pingyi Chen, Zhongyi Shui, Chenglu Zhu, and Lin Yang. Rethinking transformer for long contextual histopathology whole slide image analysis. *Advances in Neural Information Processing Systems*, 37:101498–101528, 2024.

Yan Li, Yifei Xing, Xiangyuan Lan, Xin Li, Haifeng Chen, and Dongmei Jiang. Alignmamba: Enhancing multimodal mamba with local and global cross-modal alignment. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24774–24784. IEEE, 2025.

Pei Liu, Luping Ji, Feng Ye, and Bo Fu. Advmil: Adversarial multiple instance learning for the survival analysis on whole-slide images. *Medical Image Analysis*, 91:103020, 2024.

Xiao Liu, Chenxu Zhang, Fuxiang Huang, Shuyin Xia, Guoyin Wang, and Lei Zhang. Vision mamba: A comprehensive survey and taxonomy. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570, 2021.

Lucas Schneider, Sara Laiouar-Pedari, Sara Kuntz, Eva Krieghoff-Henning, Achim Hekler, Jakob N Kather, Timo Gaiser, Stefan Froehling, and Titus J Brinker. Integration of deep learning-based image analysis and genomic data in cancer pathology: A systematic review. *European Journal of Cancer*, 160:80–91, 2022.

Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer-based correlated multiple instance learning for whole slide image classification. *Advances in neural information processing systems*, 34:2136–2147, 2021.

Andrew H Song, Guillaume Jaume, Drew FK Williamson, Ming Y Lu, Anurag Vaidya, Tiffany R Miller, and Faisal Mahmood. Artificial intelligence for digital and computational pathology. *Nature Reviews Bioengineering*, 1(12):930–949, 2023.

Suixue Wang, Xiangjun Hu, and Qingchen Zhang. Hc-mae: Hierarchical cross-attention masked autoencoder integrating histopathological images and multi-omics for cancer survival prediction. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 642–647. IEEE, 2023.

Suixue Wang, Huiyuan Lai, Shuling Wang, and Qingchen Zhang. Contramae: Contrastive alignment masked autoencoder framework for cancer survival prediction. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 2621–2626. IEEE, 2024a.

Suixue Wang, Zhigao Zheng, Xinghong Wang, Qingchen Zhang, and Zhuo Liu. A cloud-edge collaboration framework for cancer survival prediction to develop medical consumer electronic devices. *IEEE Transactions on Consumer Electronics*, 70(3):5251–5258, 2024b.

Suixue Wang, Shilin Zhang, Huiyuan Lai, Weiliang Huo, and Qingchen Zhang. Pomp: Pathology-omics multimodal pre-training framework for cancer survival prediction. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 7813–7821, 2025.

Zhikang Wang, Jiani Ma, Qian Gao, Chris Bain, Seiya Imoto, Pietro Lio, Hongmin Cai, Hao Chen, and Jiangning Song. Dual-stream multi-dependency graph neural network enables precise cancer survival analysis. *Medical Image Analysis*, 97:103252, 2024c.

Xinyu Xie, Yawen Cui, Tao Tan, Xubin Zheng, and Zitong Yu. Fusionmamba: Dynamic feature enhancement for multimodal image fusion with mamba. *Visual Intelligence*, 2(1):37, 2024.

Shu Yang, Yihui Wang, and Hao Chen. Mambamil: Enhancing long sequence modeling with sequence reordering in computational pathology. In *International Conference on Medical Image Computingg and Computer-Assisted Intervention*, pp. 296–306. Springer, 2024.

Jiawen Yao, Xinliang Zhu, Jitendra Jonnagaddala, Nicholas Hawkins, and Junzhou Huang. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Medical Image Analysis*, 65:101789, 2020.

Fengtao Zhou and Hao Chen. Cross-modal translation and alignment for survival analysis. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21428–21437. IEEE, 2023.