# Memory-Augmented Incomplete Multimodal Survival Prediction via Cross-Slide and Gene-Attentive Hypergraph Learning

Mingcheng Qu[1], Guang Yang[1], Donglin Di[2], Yue Gao[2], Tonghua Su[1], Yang Song[3], and Lei Fan[3]✉

[1]Faculty of Computing, Harbin Institute of Technology
[2]School of Software, Tsinghua University
[3]School of Computer Science and Engineering, UNSW Sydney
lei.fan1@unsw.edu.au

**Abstract.** Multimodal pathology-genomic analysis is critical for cancer survival prediction. However, existing approaches predominantly integrate formalin-fixed paraffin-embedded (FFPE) slides with genomic data, while neglecting the availability of other preservation slides, such as Fresh Froze (FF) slides. Moreover, as the high-resolution spatial nature of pathology data tends to dominate the cross-modality fusion process, it hinders effective multimodal fusion and leads to modality imbalance challenges between pathology and genomics. These methods also typically require complete data modalities, limiting their clinical applicability with incomplete modalities, such as missing either pathology or genomic data. In this paper, we propose a multimodal survival prediction framework that leverages hypergraph learning to effectively integrate multi-WSI information and cross-modality interactions between pathology slides and genomics data while addressing modality imbalance. In addition, we introduce a memory mechanism that stores previously learned paired pathology-genomic features and dynamically compensates for incomplete modalities. Experiments on five TCGA datasets demonstrate that our model outperforms advanced methods by over 2.3% in C-Index. Under incomplete modality scenarios, our approach surpasses pathology-only (3.3%) and gene-only models (7.9%). Code: https://github.com/MCPathology/M2Surv

**Keywords:** Multi-modality · Survival Analysis · Incomplete Modality

## 1 Introduction

Multimodal survival prediction, integrating Whole Slide Images (WSIs) with genomic profiles, offers great potential for advancing precision oncology [18,31]. This integration leverages the complementary strengths: WSIs capture cellular morphology and tumor micro-environment [29,23], while genomic profiles identify key driver mutations and define molecular subtypes [19,20].

✉ Corresponding author

Generally, two prevalent methods are used for WSI preparation: Formalin-Fixed Paraffin-Embedded (FFPE) and Fresh Frozen (FF). Specifically, FFPE slides are widely used due to their high-quality morphological preservation, while FF slides offer better nucleic acids and proteins but are more prone to artifacts and structural degradation. For patients with multiple slides, previous studies [3,28,30,12] typically aggregated features from all patches across different slides, overlooking the **heterogeneity** in staining style within WSIs [7].

On the other hand, these models employ mid- or late-feature fusion strategies for multimodal integration of pathology and genomics, achieving better performance compared to unimodal approaches [3,25,12]. However, they face challenges related to **modality imbalance**, as WSIs contain thousands of patches while only a few hundred genes are identified for common cancers [21]. This imbalance leads to the pathology modality dominating the fusion process, particularly when using cross-attention mechanisms [3,12]. Incorporating slides of both FFPE and FF types would further exacerbate this issue. Furthermore, in practice, technical and financial constraints often result in insufficient tissue samples and sequencing errors [6], leading to **incomplete modalities**, such as missing genomic data or pathology WSIs. However, the effectiveness of multimodal fusion strategy (*e.g.*, cross-attention) depends heavily on complete correlations between modalities, which limits their clinical applicability and poses deployment challenges in real-world settings.

In this paper, we propose M$^2$Surv, a **M**emory-augmented Incomplete **M**ultimodal **Sur**vival Prediction Framework, consisting of three stages: feature extraction, multi-slide hypergraph, and gene-attentive hypergraph. Specifically, the multi-slide hypergraph represents multiple pathology slides by treating patches as nodes, and first constructs intra-WSI hyperedges within individual slides based on spatial interactions, then progressively aggregates information across multiple slides through inter-WSI hyperedges, capturing morphological variations and histological patterns at different levels. To mitigate the pathology-genomics imbalance, the gene-attentive hypergraph establishes dense cross-modal connections by explicitly linking each gene group to all pathology features. By doing this, the importance of genomic features is reinforced, ensuring a more balanced contribution to the fusion process. Additionally, a memory bank is introduced to store paired pathology-genomic features during training, allowing the retrieval of relevant features to compensate for incomplete modalities during inference, ensuring reliable predictions even with incomplete data.

Our contributions are summarized as: A multimodal framework, M$^2$Surv, is proposed to integrate multiple pathology slides and genomic data while addressing the pathology-genomics imbalance through hypergraph learning. A memory bank is implemented with few computational consumption yet effectively compensates for missing modality. Extensive experiments demonstrate the superiority of our model, achieving advanced performance across five TCGA datasets.

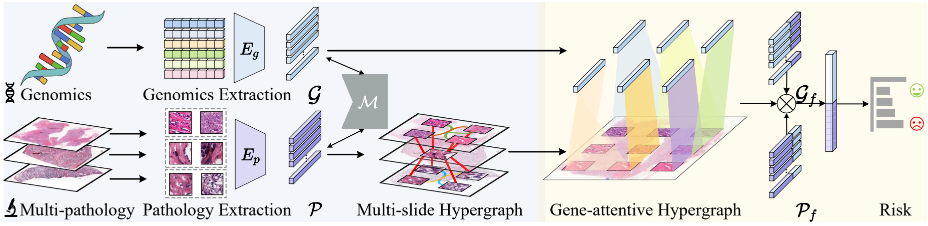

**Fig. 1. Overview of M²Surv.** It includes *feature extraction* to process multiple pathology slides and genomics, *multi-slide hypergraph* to capture both intra-slide and inter-slide feature representations, and *gene-attentive hypergraph* to establish dense connections between each gene group and all pathology patches. *A memory bank* is incorporated to store paired pathology-genomic features during training and retrieve similar features during inference, mitigating missing modality challenges.

## 2 Method

### 2.1 Overview

**Preliminary.** Given a cohort $\mathbb{X} = \{\mathrm{X}_1, \ldots, \mathrm{X}_n\}$ of $n$ subjects, each subject $\mathrm{X}_i = \{\mathrm{H}_i, y_i\}$ consists of pathology–genomics features $\mathrm{H}_i = \{\mathcal{P}_i, \mathcal{G}_i\}$ and survival information $y_i = \{c_i, t_i\}$. Here, $\mathcal{P}_i = \{\mathrm{P}_{i1}, \ldots, \mathrm{P}_{iK}\}$ includes $K$ FF and FFPE slides, $\mathcal{G}_i$ represents the genomic profiles, $c_i \in \{0, 1\}$ denotes the event status ($c_i = 0$ indicates event occurrence), and $t_i$ is the overall survival time. The goal is to estimate the hazard function $\phi_h(t)$, which predicts the instantaneous incidence event rate at time $t$, while training a model $\mathcal{F}$ to predict the probability of survival beyond $t$ using the survival function $\phi_s(t)$. The model is optimized using the negative log-likelihood losses [26], defined as:

$$\mathcal{L}_{\mathcal{S}} = \textstyle\sum_{i=1}^{n} (1 - c_i) \log \phi_h(t_i|\mathrm{H_i}) + c_i \log \phi_s(t_i|\mathrm{H_i}) + (1 - c_i) \log \phi_s(t_i - 1|\mathrm{H_i}). \quad (1)$$

**Our Framework.** M²Surv comprises feature extraction, multi-slide hypergraph, and gene-attentive hypergraph (see Fig. 1). Paired pathology $\mathcal{P}$ and genomics $\mathcal{G}$ features are extracted using respective encoders. The multi-slide hypergraph first builds intra-WSI graphs for each slide based on spatial interactions, followed by an inter-WSI hypergraph to capture structural relationships across multiple slides. The gene-attentive hypergraph constructs dense connections using gene features to guide and refine $\mathcal{P}_h$, generating integrated features $\mathcal{P}_f$ and $\mathcal{G}_f$ for final risk prediction. To handle incomplete modalities, a memory bank $\mathcal{M}$ is introduced to store paired pathology-genomic features $\mathrm{M}_p$ and $\mathrm{M}_g$ during training, allowing the model to retrieve and approximate the similar features from $\mathcal{M}$ to compensate for missing information during inference.

**Feature Extraction.** Following previous studies [16,12,28], each WSI $\mathrm{P}_k$ is partitioned and randomly selected into $N_k = 4096$ patches of $256 \times 256$ pixel, at 20× magnification. A pretrained encoder (*e.g.,* ResNet50) extracts $d$-dimensional features from these patches, representing the WSI as $\mathrm{P}_k \in \mathbb{R}^{N_k \times d} = \{p_1, \ldots, p_{N_k}\}$, where each patch $p_k$ has spatial coordinates $\zeta_{p_k} = (x_k, y_k)$. The multi-pathology feature set $\mathcal{P} = \{\mathrm{P}_1, \dots, \mathrm{P}_K\}$ includes multiple slide features from the same patient. For genomic data like RNA-seq, CNV, and SNV, we adopt the feature selection method [3] by grouping them into $W = 6$ functional groups: Tumor Suppression, Oncogenesis, Kinases, Cellular Differentiation, Transcription, and Cytokines. Each category is encoded using a genomic encoder (*i.e.*, a multilayer perceptron, MLP) to produce genomic features $\mathcal{G} \in \mathbb{R}^{W \times d} = \{g_1, \dots, g_W\}$.

### 2.2 Multi-slide Hypergraph

Considering the distinct staining styles and biopsy tissue variations in multiple slides, we adopt a two-stage strategy: intra-slide to handle style differences within individual slides, followed by inter-slide integration to ensure the aggregation of multi-slide information. Specifically, we leverage widely used hypergraphs [4,5,13,27] to construct **intra-slide topological hyperedges** for each WSI while establishing **inter-slide structural hyperedges** across different slides.

For a WSI $\mathrm{P}_k$, each patch $p$ is treated as a vertex, and intra-slide hyperedges are formed by grouping each patch with its neighboring patches based on Euclidean distance. Given a patch $p_k$, its neighbors are determined as:

$$\mathcal{N}_T(p_k) = \{p_j \mid \zeta_{p_j} - \zeta_{p_k}\|_2 \leq \delta\}, \tag{2}$$

where $\zeta_{p_j}$ and $\zeta_{p_k}$ denote the coordinates of patches $p_j$ and $p_k$ respectively, and $\delta$ is a distance threshold. This yields topological-based hyperedges: $\mathcal{E}_T^{(k)} = \{\{p_k, p_{j1}, p_{j2}, \dots\} \mid \forall p_j \in N_T(p_k)\}$. The combined intra-slide hyperedges across $K$ WSIs are then represented as $\mathcal{E}_T = \{\mathcal{E}_T^{(1)}, \dots, \mathcal{E}_T^{(K)}\}$. By capturing the topological relationships between patches, this approach effectively encodes spatial structures within each WSI, thereby preserving the inherent tissue morphology.

For multi-pathology $\mathcal{P}$, all patches are treated as nodes, and inter-slide structural hyperedges using feature similarity. The neighbors of $p_k$ are identified as:

$$\mathcal{N}_F(p_k) = \{p_j \mid \mathrm{sim}(p_k, p_j) \geq \alpha\}, \tag{3}$$

where $\mathrm{sim}(\cdot, \cdot)$ represents the cosine similarity function, and $\alpha$ is a similarity threshold. This builds structural hyperedges: $\mathcal{E}_F = \{\{p_k, p_{j_1}, p_{j_2}, \dots\} \mid \forall\ p_j \in \mathcal{N}_F(p_k)\}$. By capturing the structural relationships between patches across multiple WSIs, it identifies shared morphological patterns and preserves consistent biological information across WSIs. Notably, $\delta$ and $\alpha$ are determined by the hyperedge construction threshold $\lambda$, which selects $\lambda - 1$ neighbors for each patch.

The final hyperedge set $\mathcal{E}_m$ is formed by merging these intra-slide and inter-slide hyperedges, where $\mathcal{E}_m = \mathcal{E}_T \cup \mathcal{E}_F$ represents their union. The constructed multi-slide hypergraph is then processed through hypergraph convolutions [8] to extract high-order feature representations $\mathcal{P}_h = \mathcal{P}^{(L)}$ after $L$ layers.

### 2.3 Gene-attentive Hypergraph

Considering the dimensional imbalance between genomic and pathology features, we aim to construct a dense gene-attentive hypergraph by establishing connections between each gene group and the multi-slide hypergraph. It is motivated by the ability of hypergraphs to link features from different modalities into interconnected clusters, effectively capturing complex cross-modal relationships within a unified structure. Specifically, all nodes in the multi-slide hypergraph and the six gene groups are treated as new nodes, with hyperedges $\mathcal{E}_g$ centered around each gene group. The neighbors are identified as:

$$\mathcal{N}_G(g_w) = \{p_j \mid \frac{\exp(\mathrm{att}(g_w, p_j))}{\sum_{k=1}^{N} \exp(\mathrm{att}(g_w, p_k))} \geq \beta\}, \tag{4}$$

where *att* represents the cross-attention score between gene group $g_w$ and patch $p_j$, and $\beta$ is the threshold (empirically set to select the top 5% of patch-gene hyperedges balancing connections and computation). This results gene-attentive hyperedges: $\mathcal{E}_G = \{\{g_w, p_{j_1}, p_{j_2}, \dots\} \mid \forall\ p_j \in \mathcal{N}_G(g_w)\}$. Then, the hypergraph convolution is employed to update the representations of all nodes and produce the refined pathology and genomic features $\mathcal{P}_f$ and $\mathcal{G}_f$.

The gene-attentive hypergraph leverages cross-attention mechanisms to define hyperedges, emphasizing cross-modal relationships and interactions. By establishing dense connections centered on gene groups, our approach enables each gene group to effectively interact with multiple pathology patches, mitigating the imbalance issue observed in previous cross-attention strategies [3,12].

### 2.4 Memory Bank

To address incomplete modalities in clinical scenarios, we introduce a memory bank $\mathcal{M}$ [1] to store paired pathology-genomic features $\{\langle\mathcal{P}, \mathcal{G}\rangle\}_{i=1}^{n}$ during training, where $n$ is the number of training samples. We employ a momentum update strategy [10] to dynamically update stored features. At the training $r$- epoch, $\mathcal{M}$ is updated: $\mathcal{M}^{(r)} \leftarrow \theta \cdot \left\langle \mathcal{P}^{(r)}, \mathcal{G}^{(r)} \right\rangle + (1-\theta) \cdot \mathcal{M}^{(r-1)}$, where $\theta \in [0, 1]$ is the momentum coefficient.

During inference, if a modality (*i.e.*, all pathology slides or genomic data) is missing, the available modality $\mathrm{M_a}$ is used to retrieve the most relevant features by computing its feature similarity with all entries in $\mathcal{M}$. The top-$\mu$ most similar entries $\{\mathrm{M}_m^{(j)}\}_{j=1}^{\mu}$ are selected and then aggregated, expressed as:

$$\hat{\mathrm{M}}_m = \sum_{j=1}^{\mu} \frac{\exp(\mathrm{sim}(\mathrm{M}_a, \mathrm{M}_a^{(j)}))}{\sum_{l=1}^{\mu} \exp(\mathrm{sim}(\mathrm{M}_a, \mathrm{M}_a^{(l)}))} \mathrm{M}_m^{(j)}, \tag{5}$$

where $\hat{\mathrm{M}}_m$ is the approximated features used to represent the missing features.

Our memory bank leverages momentum update retrieval for collective analysis of historical training data. Essentially, it integrates multimodal information from previously learned multiple samples, effectively addressing incomplete modalities while maintaining efficiency.

**Table 1. Comparison of our model with advanced methods on five datasets.** C-Indexes (Mean ± STD) are reported based on 5-fold cross-validation.

| | Model | BLCA | BRCA | CO-READ | HNSC | STAD | Mean |
|---|---|---|---|---|---|---|---|
| **Pathology** | ABMIL [11] | 0.624 ± 0.059 | 0.672 ± 0.051 | 0.730 ± 0.151 | 0.624 ± 0.042 | 0.636 ± 0.043 | 0.657 |
| | AMISL [26] | 0.627 ± 0.032 | 0.681 ± 0.036 | 0.710 ± 0.091 | 0.607 ± 0.048 | 0.553 ± 0.012 | 0.636 |
| | TranMIL [22] | 0.617 ± 0.045 | 0.663 ± 0.053 | 0.747 ± 0.151 | 0.619 ± 0.062 | 0.660 ± 0.072 | 0.661 |
| | CLAM-MB [16] | 0.623 ± 0.045 | 0.696 ± 0.098 | 0.721 ± 0.159 | **0.648 ± 0.050** | 0.620 ± 0.034 | 0.662 |
| | M$^2$Surv (**Ours**) | **0.646 ± 0.022** | **0.735 ± 0.056** | **0.749 ± 0.045** | 0.612 ± 0.004 | **0.677 ± 0.048** | **0.684** |
| **Genomic** | MLP [9] | 0.530 ± 0.070 | 0.622 ± 0.079 | 0.712 ± 0.114 | 0.520 ± 0.064 | 0.497 ± 0.031 | 0.576 |
| | SNN [15] | 0.521 ± 0.070 | 0.621 ± 0.073 | 0.711 ± 0.162 | 0.514 ± 0.076 | 0.485 ± 0.047 | 0.570 |
| | SNNTrans [15] | 0.583 ± 0.060 | 0.679 ± 0.053 | **0.739 ± 0.124** | 0.570 ± 0.035 | 0.547 ± 0.041 | 0.622 |
| | M$^2$Surv (**Ours**) | **0.593 ± 0.065** | **0.696 ± 0.043** | 0.701 ± 0.087 | **0.659 ± 0.058** | **0.704 ± 0.082** | **0.671** |
| **Multimodal** | SNN+CLAM | 0.625 ± 0.060 | 0.699 ± 0.064 | 0.716 ± 0.016 | 0.638 ± 0.066 | 0.629 ± 0.065 | 0.661 |
| | Porpoise [2] | 0.617 ± 0.056 | 0.668 ± 0.070 | 0.738 ± 0.151 | 0.614 ± 0.058 | 0.660 ± 0.106 | 0.659 |
| | MCAT [3] | 0.640 ± 0.076 | 0.685 ± 0.109 | 0.724 ± 0.137 | 0.564 ± 0.084 | 0.625 ± 0.118 | 0.648 |
| | MOTCat [25] | 0.659 ± 0.069 | 0.727 ± 0.027 | 0.742 ± 0.124 | 0.656 ± 0.041 | 0.621 ± 0.065 | 0.681 |
| | CMTA [30] | 0.670 ± 0.030 | 0.691 ± 0.037 | 0.704 ± 0.117 | 0.562 ± 0.086 | 0.592 ± 0.014 | 0.644 |
| | SurvPath [12] | 0.635 ± 0.026 | 0.679 ± 0.077 | 0.731 ± 0.124 | 0.617 ± 0.058 | 0.620 ± 0.044 | 0.656 |
| | PIBD [28] | 0.651 ± 0.092 | 0.712 ± 0.048 | **0.786 ± 0.134** | 0.607 ± 0.059 | 0.668 ± 0.055 | 0.685 |
| | M$^2$Surv (**Ours**) | **0.671 ± 0.039** | **0.744 ± 0.091** | 0.757 ± 0.080 | **0.661 ± 0.012** | **0.673 ± 0.028** | **0.701** |

## 3 Experiments

**Datasets and Experimental Settings.** Following previous studies [12,28], we evaluated our models on five The Cancer Genome Atlas (TCGA) datasets: Bladder Urothelial Carcinoma (BLCA) (n=384), Breast Invasive Carcinoma (BRCA) (n=968), Colon and Rectum Adenocarcinoma (CO-READ) (n=298), Head and Neck Squamous Cell Carcinoma (HNSC) (n=392), and Stomach Adenocarcinoma (STAD) (n=317). We followed the previous dataset settings [3,12,28], and employed the Adam optimizer with a learning rate of $1 \times 10^{-4}$ and a weight decay of $1 \times 10^{-5}$, training 30 epochs. Concordance Index (C-Index) was used as the metric. For each dataset, We performed 5-fold cross-validation with a 4:1 train-val split, reporting results as the mean ± standard deviation (STD).

### 3.1 Comparisons with Advanced Methods

We compared our models in three settings: pathology-only (ABMIL, AMISL, TransMIL, and CLAM), genomic-only (MLP, SNN, and SNNTrans), and multimodal (Porpoise, MCAT, MOTCat, CMTA, SurvPath, and PIBD). Results for unimodal methods, SNN+CLAM, and Porpoise were quoted from previous study [28], while others were reproduced using their released code (see Table 1).

Our model achieved a mean C-index of 0.701, outperforming all unimodal and multimodal methods on nearly all datasets. In modality-missing scenarios, it exhibited superior performance with a mean C-Index of 0.684 (pathology-only), surpassing TransMIL (0.661) and CLAM-MB (0.662). Similarly, it maintained a score of 0.671, outperforming gene-only models such as SNNTrans (0.622). These results demonstrate the effectiveness of our framework in integrating pathology and genomic data while handling incomplete modalities.

**Table 2. Ablation study on multi-slide hypergraph and gene-attentive hypergraph**. We evaluated pathology aggregator (*Agg*), slide type (HGNN$^-$ for FFPE only), hyperedge types, and construction threshold $\lambda$ for multi-slide hypergraph, and multimodal feature fusion (*Fuse*) and cross-modal edge construction methods (HGNN$^*$ for random edge instead of attention score) for gene-attentive hypergraph.

| | *Agg* | Slides | $\lambda$ | *Fuse* | BLCA | BRCA | CO-READ | HNSC | STAD | Mean |
|---|---|---|---|---|---|---|---|---|---|---|
| **Multi-slide hypergraph** | MLP | *Multi* | 9 | HGNN | 0.644 ± 0.064 | 0.708 ± 0.039 | 0.750 ± 0.046 | 0.620 ± 0.043 | 0.638 ± 0.056 | 0.672 |
| | ABMIL | *Multi* | 9 | HGNN | 0.645 ± 0.034 | 0.715 ± 0.101 | 0.699 ± 0.049 | 0.632 ± 0.037 | 0.661 ± 0.031 | 0.670 |
| | TransMIL | *Multi* | 9 | HGNN | 0.634 ± 0.045 | 0.718 ± 0.083 | 0.743 ± 0.109 | 0.623 ± 0.038 | 0.661 ± 0.056 | 0.676 |
| | GAT | *Multi* | 9 | HGNN | 0.648 ± 0.039 | 0.733 ± 0.116 | 0.730 ± 0.136 | 0.637 ± 0.040 | 0.656 ± 0.094 | 0.681 |
| | GCN | *Multi* | 9 | HGNN | 0.644 ± 0.038 | 0.701 ± 0.053 | 0.755 ± 0.086 | 0.621 ± 0.020 | 0.667 ± 0.044 | 0.678 |
| | HGNN$^-$ | intra. | 9 | HGNN | 0.668 ± 0.045 | 0.733 ± 0.062 | **0.767 ± 0.095** | 0.639 ± 0.056 | 0.671 ± 0.067 | 0.696 |
| | HGNN | intra. | 9 | HGNN | 0.645 ± 0.041 | 0.684 ± 0.065 | 0.706 ± 0.122 | 0.631 ± 0.021 | 0.661 ± 0.066 | 0.665 |
| | HGNN | inter. | 9 | HGNN | 0.632 ± 0.024 | 0.680 ± 0.103 | 0.742 ± 0.062 | 0.627 ± 0.039 | 0.639 ± 0.079 | 0.664 |
| | HGNN | *Multi* | 5 | HGNN | 0.649 ± 0.032 | 0.695 ± 0.048 | 0.689 ± 0.067 | 0.641 ± 0.082 | **0.706 ± 0.088** | 0.676 |
| | HGNN | *Multi* | 25 | HGNN | 0.640 ± 0.028 | 0.734 ± 0.061 | 0.709 ± 0.080 | 0.620 ± 0.040 | 0.660 ± 0.067 | 0.673 |
| **Gene-attn** | HGNN | *Multi* | 9 | Concat | 0.628 ± 0.049 | 0.720 ± 0.035 | 0.708 ± 0.026 | 0.590 ± 0.065 | 0.673 ± 0.048 | 0.664 |
| | HGNN | *Multi* | 9 | Co-Attn | 0.669 ± 0.030 | 0.677 ± 0.035 | 0.713 ± 0.087 | 0.632 ± 0.033 | 0.655 ± 0.048 | 0.669 |
| | HGNN | *Multi* | 9 | GAT | 0.623 ± 0.039 | 0.731 ± 0.089 | 0.717 ± 0.126 | 0.642 ± 0.070 | 0.651 ± 0.059 | 0.673 |
| | HGNN | *Multi* | 9 | GCN | 0.646 ± 0.037 | 0.732 ± 0.095 | 0.692 ± 0.097 | 0.607 ± 0.059 | 0.703 ± 0.091 | 0.676 |
| | HGNN | *Multi* | 9 | HGNN$^*$ | 0.601 ± 0.011 | 0.702 ± 0.066 | 0.686 ± 0.104 | 0.629 ± 0.063 | 0.654 ± 0.030 | 0.654 |
| | HGNN | *Multi* | 9 | HGNN | **0.671 ± 0.039** | **0.744 ± 0.091** | 0.757 ± 0.080 | **0.661 ± 0.012** | 0.673 ± 0.028 | **0.701** |

### 3.2 Ablation Studies

**Multi-slide hypergraph.** We evaluated various pathology aggregators (*Agg*), slide types, hyperedge types, and hyperedge construction thresholds $\lambda$ (see Table 2). When using *Agg* such as MLP (0.672), ABMIL [11] (0.670), and TransMIL [22] (0.676), incorporating GAT [24], GCN [14] and HGNN improved the average C-Index to 0.681, 0.698 and 0.701 respectively. highlighting the effectiveness of graph-based models in enhancing feature aggregation. Using only FFPE (HGNN$^-$) yielded a score of 0.696, which increased to 0.701 with multi-slide integration, demonstrating the benefit of incorporating multiple slides.

When exploring different slide types, using only intra-slide topological hyperedges or inter-slide structural hyperedges achieved a mean C-Indexes of 0.665 and 0.664 respectively, indicating that both connections contribute to model performance. Moreover, varying HGNN construction thresholds (5, 9, and 25) produced scores of 0.676, 0.701, and 0.673, respectively. This suggests that an optimal threshold balances information retention, preventing homogenization at ($\lambda = 25$) and avoiding oversimplification at ($\lambda = 5$).

**Gene-attentive hypergraph.** We evaluated multimodal fusion and hyperedge construction methods (see Table 2). Compared to direct concatenation (0.664) and cross-attention (0.669), gene-attentive GAT and GCN improved the C-Index to 0.673 and 0.676, respectively. HGNN with attention-based edge construction achieved a score of 0.701 better than random edge construction (0.654), confirming the effectiveness of hypergraph integration with attention mechanisms in capturing cross-modal interactions and mitigating modality imbalance.

**Memory bank.** We evaluated the time consumption, top-$\mu$, and generalizability of our memory bank across five datasets (see Fig. 2). Specifically, using

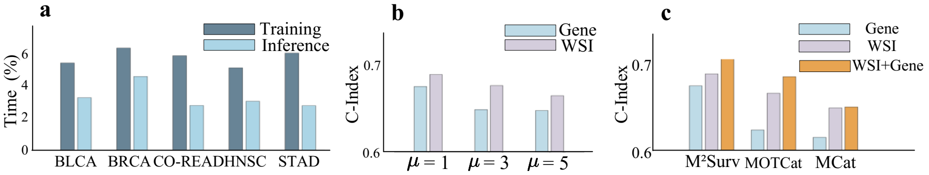

**Fig. 2. a)**, Extra time consumption for memory bank during training and inference. **b)**, Performance with varying retrieval top-$\mu$. **c)**, Performance with incomplete modality using the memory bank across models (M²Surv, MCat,and MOTCat).

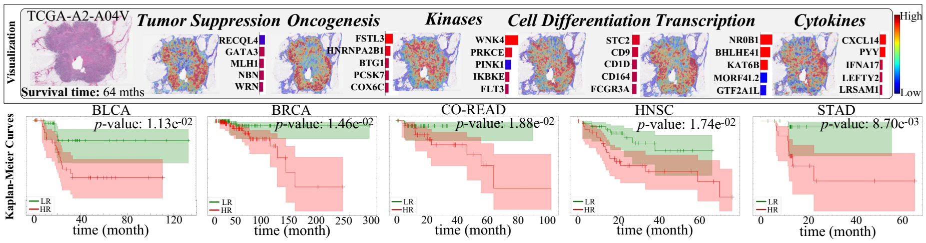

**Fig. 3. Visualization for BRCA cases:** The heatmap generated from cross-attention scores, and the top five most influential genes are highlighted by gradient integral. **Kaplan-Meier curves (Bottom)** shows significant survival stratification (p<0.05 across all datasets) between high/low-risk groups (median split).

the memory bank across various datasets resulted in an approximately 4.5% increase in overall time consumption and a 2.5% in inference time. When selecting the top $\mu$ relevant features, the best average performance was achieved at $\mu = 1$, indicating that utilizing only the most relevant stored feature is effective in mitigating missing modalities. Incorporating the memory bank into MCat [3] and MOTCat [25] enabled these models to perform effectively even with incomplete modalities while maintaining performance comparable to full settings, showing the adaptability of the memory bank in enhancing multimodal methods.

### 3.3 Visualization

The attention score visualizations highlight specific regions targeted by different gene groups, quantitatively assessing the gene-attentive hypergraph in capturing cross-modal interactions (see Fig. 3). For example, it reveals individual gene influences, with NR0B1 exhibiting a high positive gradient, indicating an enhancing role in certain pathological conditions. To validate the discriminative capacity of our model, we performed Kaplan-Meier analysis and log-rank test [17] by stratifying patients into high- and low-risk groups based on the predicted median risk scores. The $p$-values below 0.05 confirmed the effectiveness of the model across all datasets.

## 4 Conclusion

We proposed M$^2$Surv, a multimodal survival prediction framework that leverages hypergraphs to model multi-slide and genomic data. We introduced a memory bank to handle missing clinical modalities. Extensive experiments demonstrated the superior performance of our framework.

**Limitations.** While incorporating more slides enhances consistency modeling, artifacts in FF slides may introduce noise into pathology features. Hypergraphs effectively capture multi-scale interaction but are less flexible than cross-attention mechanisms in dynamically adjusting modality-specific contribution weights for fine-grained feature alignment. Moreover, the retrieval efficacy of memory banks depends on the coverage of training data, which may struggle to capture rare pathology-genomics associations in historical records.

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

## References

1. Alonso, I., Sabater, A., Ferstl, D., Montesano, L., Murillo, A.C.: Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank. In: ICCV. pp. 8219–8228 (2021)
2. Chen, R.J., Lu, M.Y., Williamson, D.F., et al.: Pan-cancer integrative histology-genomic analysis via multimodal deep learning. Cancer Cell **40**(8), 865–878 (2022)
3. Chen, R.J., Lu, M.Y., et al.: Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In: ICCV. pp. 3995–4005 (2021). `https://doi.org/10.1109/ICCV48922.2021.00398`
4. Di, D., Zhang, J., Lei, F., Tian, Q., Gao, Y.: Big-hypergraph factorization neural network for survival prediction from whole slide image. IEEE Transactions on Image Processing **31**, 1149–1160 (2022)
5. Di, D., Zou, C., Feng, Y., et al.: Generating hypergraph-based high-order representations of whole-slide histopathological images for survival prediction. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(5), 5800–5815 (2022)
6. Dorent, R., Joutard, S., Modat, M., Ourselin, S., Vercauteren, T.K.M.: Hetero-modal variational encoder-decoder for joint modality completion and segmentation. In: MICCAI (2019), `https://api.semanticscholar.org/CorpusID:198897112`
7. Fan, L., Sowmya, A., Meijering, E., Song, Y.: Cancer survival prediction from whole slide images with self-supervised learning and slide consistency. IEEE Transactions on Medical Imaging **42**(5), 1401–1412 (2022)
8. Feng, Y., You, H., Zhang, Z., Ji, R., Gao, Y.: Hypergraph neural networks. In: AAAI. vol. 33, pp. 3558–3565 (2019)
9. Haykin, S.: Neural networks: a comprehensive foundation. Prentice Hall PTR (1998)
10. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR. pp. 9729–9738 (2020)
11. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: ICML. pp. 2127–2136. PMLR (2018)

12. 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: CVPR. pp. 11579–11590 (2024)
13. Jing, W., Wang, J., Di, D., et al.: Multi-modal hypergraph contrastive learning for medical image segmentation. Pattern Recognition **165**, 111544 (2025)
14. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
15. Klambauer, G., Unterthiner, T., Mayr, A., Hochreiter, S.: Self-normalizing neural networks. NeurIPS **30** (2017)
16. Lu, M.Y., Williamson, D.F., Chen, T.Y., et al.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature biomedical engineering **5**(6), 555–570 (2021)
17. Mantel, N., et al.: Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemother Rep **50**(3), 163–170 (1966)
18. Nunes, L., Li, F., Wu, M., et al.: Prognostic genome and transcriptome signatures in colorectal cancers. Nature **633**(8028), 137–146 (2024)
19. Qu, M., Wu, Y., Di, D., Su, A., Su, T., Song, Y., Fan, L.: Boundary-guided learning for gene expression prediction in spatial transcriptomics. In: BIBM. pp. 445–450. IEEE (2024)
20. Qu, M., Yang, G., Su, T., Gao, Y., Song, Y., Fan, L., et al.: Multimodal cancer survival analysis via hypergraph learning with cross-modality rebalance. arXiv preprint arXiv:2505.11997 (2025)
21. Raser, J.M., O'shea, E.K.: Noise in gene expression: origins, consequences, and control. Science **309**(5743), 2010–2013 (2005)
22. 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. NeurIPS **34**, 2136–2147 (2021)
23. Tang, Q., Fan, L., Pagnucco, M., Song, Y.: Prototype-based image prompting for weakly supervised histopathological image segmentation. In: CVPR. pp. 30271–30280 (2025)
24. Velickovic, P., Cucurull, G., Casanova, A., et al.: Graph attention networks. stat **1050**(20), 10–48550 (2017)
25. Xu, Y., Chen, H.: Multimodal optimal transport-based co-attention transformer with global structure consistency for survival prediction. In: ICCV. pp. 21241–21251 (2023)
26. Yao, J., Zhu, X., et al.: Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. Medical Image Analysis **65**, 101789 (2020)
27. Yi, K., Chen, J., Wang, Y.G., et al.: Approximate equivariance so (3) needlet convolution. In: Topological, Algebraic and Geometric Learning Workshops 2022. pp. 189–198. PMLR (2022)
28. Zhang, Y., Xu, Y., Chen, J., Xie, F., Chen, H.: Prototypical information bottlenecking and disentangling for multimodal cancer survival prediction. In: ICLR (2024)
29. Zhang, Z., Chen, P., McGough, M., et al.: Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. Nature Machine Intelligence **1**(5), 236–245 (2019)
30. Zhou, F., Chen, H.: Cross-modal translation and alignment for survival analysis. In: ICCV. pp. 21485–21494 (2023)
31. Zhu, Z., Fan, L., Pagnucco, M., Song, Y.: Interpretable image classification via non-parametric part prototype learning. In: CVPR. pp. 9762–9771 (2025)