# ProtoDiffSurv: Differential Prototype Learning for Weakly Supervised Survival Analysis

Jiahao Li[1] Shu Zhang[1*]

College of Computer and Data Science, Fuzhou University, China
zhangshu@fzu.edu.cn

**Abstract.** Survival prediction from whole-slide histopathology images (WSIs) and genomic profiles is pivotal for cancer prognosis but faces significant challenges. Under weak supervision, survival-related features are sparse: only a few patches within a gigapixel WSI are truly risk-relevant, while most are irrelevant. Reliable extraction of prognostically relevant features is already challenging under such imbalance; missing modalities further compound this difficulty. Existing methods usually rely on imputation or reconstruction, which can introduce noise. As a result, models may focus on redundant synthesized patterns instead of survival-critical cues, potentially obscuring survival-critical cues. To address these issues, we propose ProtoDiffSurv, a dual-stage framework that introduces modality-specific prototype banks as stable semantic anchors. Guided by these prototypes, the Prototype-Guided Patch Enhancement module identifies and up-weights risk-relevant regions, increasing their influence during training. To reduce imputation noise, we introduce Cross-Modal Differential Attention, which removes shared noise-like components via a differential operation and emphasizes modality-specific risk patterns. The resulting differential features serve as a strong semantic complement, yielding enhanced risk features. Extensive experiments on five TCGA benchmarks demonstrate that ProtoDiffSurv achieves state-of-the-art performance, outperforming existing methods in both multimodal and missing-modality settings. The source code is available at https://github.com/LJHao1208/ProtoDiffSurv.

**Keywords:** Survival Analysis · Multimodal Learning · Differential Attention · Prototype Purification · Missing Modality

## 1 Introduction

Survival analysis [4] on whole-slide images (WSIs) has attracted increasing attention due to its potential to capture rich morphological patterns associated with tumor progression and patient prognosis. Meanwhile, advances in deep learning have further propelled this field, enabling models to learn representations that are relevant for prognosis. Due to their gigapixel resolution, WSIs are divided into patches, while survival labels are available only at the patient or slide level,

* Corresponding author.

making survival analysis inherently weakly supervised. In addition, survival-discriminative signals are sparse across patches, further increasing the difficulty of learning under weak supervision.

Most WSI-based survival analysis methods adopt a multiple instance learning (MIL) paradigm, aggregating patch representations into a slide/patient-level feature via attention mechanisms [7, 15, 17–19, 1, 5]. However, sparse prognostic patches may be overwhelmed by irrelevant tissue. To capture their spatial relationships, GNNs have been introduced [12, 19], but their high computational cost at the WSI scale limits long-range interaction modeling. Recent work has begun to explore prototype-based learning [14, 26]. Nevertheless, Prototype-based methods mainly rely on similarity matching, but in gigapixel WSIs, abundant background often looks similar to informative regions, thereby blurring true prognostic signals.

To improve survival prediction, genomic data is integrated with WSIs to provide molecular prognostic cues beyond morphology, thereby enabling more robust, patient-specific risk estimation [3, 8, 21, 27]. However, in clinical practice, missing modalities are common [16]. Imputation/reconstruction [11] and distillation [2, 13, 22, 6] are the two most common strategies for handling missing modalities. Reconstruction and distillation emphasize matching the missing modality or teacher. But this kind of similarity does not guarantee better risk prediction and may capture non-prognostic details.

Unlike existing methods that frame missing modalities as a reconstruction or imputation problem, we reformulate multimodal survival prediction as semantic risk denoising. We argue that modality-shared confounders, rather than missing features themselves, are the main barrier to reliable prognosis. Accordingly, we propose ProtoDiffSurv, a Differential Prototype Rectification framework that uses dual-stage refinement and differential prototype retrieval to cancel shared confounders while preserving prognostically discriminative residuals. This design distinguishes ProtoDiffSurv from conventional imputation- and prototype-based approaches, improving prognostic accuracy and robustness.

**Contributions of our work**: 1) We propose a framework, ProtoDiffSurv, that extracts discriminative risk representations via a dual-stage refinement process. 2) To mitigate sparse risk signals in WSIs, we propose a Prototype-Guided Patch Enhancement module. It serves as a spatial filter, suppressing background noise by reweighting patches using a learnable prototype bank. 3) We propose a Cross-Modal Differential Attention mechanism to extract discriminative risk semantics by dynamically filtering the "common noise". It adaptively subtracts a noise-focused attention map from a signal-focused one, removing common-noise components from the prototype banks.

## 2 Methodology

As illustrated in Fig. 1, we propose ProtoDiffSurv, a prototype-driven framework designed to extract discriminative risk representations via a dual-stage refinement process. First, to tackle sparse prognostic signals, raw WSI patches

are fed into a Prototype-Guided Patch Enhancement module to attenuate background artifacts. Second, to reduce common noise that can mislead imputation, Cross-Modal Differential Attention is used to refine the features. It detects modality-shared background patterns from the prototype banks and subtracts them, thereby emphasizing prognostic, modality-specific cues. Finally, using a residual connection to merge the refined and original features, reducing noise, and supporting robust inference with either complete modalities or missing ones.

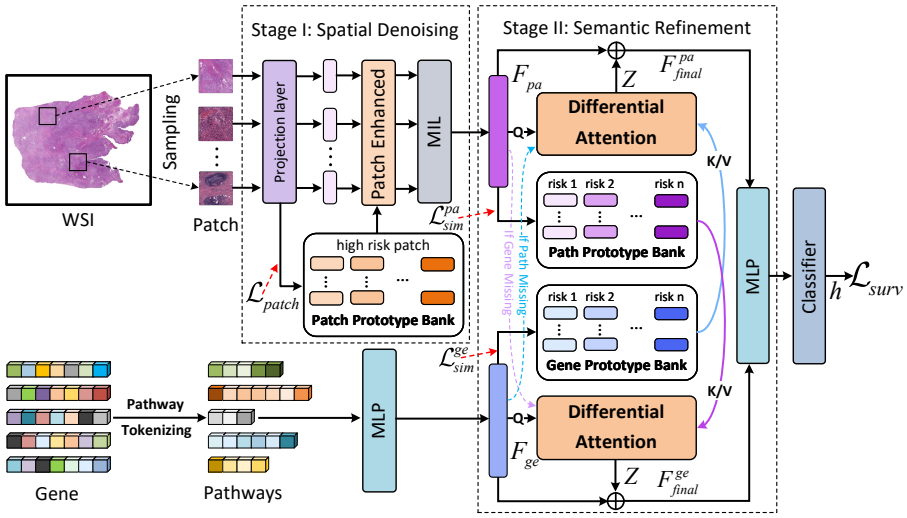


**Fig. 1.** Overall architecture of ProtoDiffSurv. The framework employs a dual-stage : (1) **Stage I (Spatial Denoising)** suppresses background in sparse WSIs via Prototype-Guided Patch Enhancement; (2) **Stage II (Semantic Refinement)** anchors risk patterns using Modality-Specific Prototype Banks; and (3) **Cross-Modal Differential Attention** explicitly cancels "common noise." Notably, dashed lines indicate the missing-modality mechanism: querying the missing modality's bank to retrieve semantics. Finally, features are fused via residuals for rectification.

### 2.1 Prototype-Guided Patch Enhancement

**Patch Projection** Each WSI is divided into non-overlapping patches. A feature extractor and projection layer map these patches into a shared embedding space:

$$\mathbf{X} = \{x_1, x_2, \ldots, x_N\}, \quad x_i \in \mathbb{R}^D, \tag{1}$$

where $N$ denotes the number of patches and $D$ is the embedding dimension.

**Patch Prototype Bank** To capture survival-relevant pathological patterns, we introduce a learnable patch prototype bank:

$$\mathcal{P}^{\text{patch}} = \{p_1, p_2, \ldots, p_K\}, \tag{2}$$

where $K = 32$ denotes the number of prototypes. This compact bank aims to aggregate risk-relevant local patterns under weak supervision. For a reliable warm start, the prototype bank is initialized via K-means clustering on training-set features and subsequently updated via backpropagation the network.

**Prototype-Guided Patch Weighting** To perform spatial denoising, we compute the similarity between each patch $x_i$ and the prototype bank. Normalized attention weights are obtained as:

$$\alpha_{i,k} = \text{Softmax}\left(\frac{\text{sim}(x_i, p_k)}{\tau}\right), \tag{3}$$

where $\tau$ is a temperature parameter. The patch score is derived from the maximum attention weight $w_i = \max \alpha_{i,k}$, and the enhanced patch representation:

$$\tilde{x}_i = w_i \cdot x_i. \tag{4}$$

This strategy amplifies patches aligning with risk prototypes while suppressing non-survival regions. Finally, the enhanced patches $\{\tilde{x}_i\}$ are aggregated into a slide-level representation $F_{pa}$ via attention-based MIL [7], serving as the denoised query for the subsequent cross-modal differential stage.

### 2.2 Modality-Specific Prototype Banks

For genomics, 1D RNA-Seq profiles are tokenized into biological pathways and projected via an MLP to obtain the initial feature $F_{ge}$. To support multimodal modeling, we establish separate prototype banks for pathology and genomics:

$$\mathcal{P}^{\text{pa}} = \{p_1, \ldots, p_M\}, \quad \mathcal{P}^{\text{ge}} = \{g_1, \ldots, g_M\}. \tag{5}$$

These larger banks ($M = 128$) encode highly heterogeneous, multi-level risk prototypes, serving as a reliable knowledge base for subsequent differential refinement.

### 2.3 Differential Prototype-Based Semantic Refinement

To effectively identify risk patterns, we introduce a Cross-Modal Differential Attention (CDA) mechanism. CDA extends the differential-attention idea of the Differential Transformer [24] to a cross-modal retrieval setting. Instead of modeling intra-sequence dependencies, CDA uses the available-modality feature as queries to retrieve semantics from a target prototype bank.

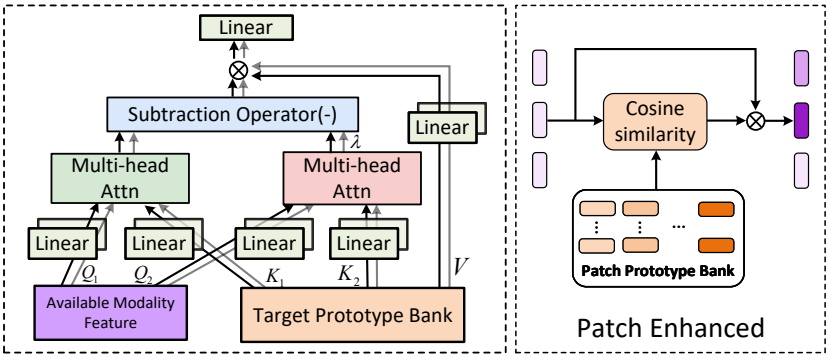


**Fig. 2.** Details of the dual-denoising mechanisms. (Left) Cross-Modal Differential Attention: Queries the Target Prototype Bank and employs a Subtraction Operator (-) to explicitly cancel "common noise," thereby extracting refinement risk semantics without imputation artifacts. (Right) Prototype-Guided Patch Enhancement: Addresses risk sparsity by filtering raw patches against the Patch Prototype Bank via Cosine similarity, highlighting high-risk regions while suppressing background noise.

As illustrated in Fig. 2 (Left), the module takes an available modality feature $\mathbf{F}$ and a target prototype bank $\mathcal{P}$. Concretely, we compute two multi-head cross-attention responses with separate $(\mathbf{Q}_1, \mathbf{K}_1)$ and $(\mathbf{Q}_2, \mathbf{K}_2)$ projections (sharing values $\mathbf{V}$), and take their difference:

$$\mathbf{A} = \text{Softmax}\left(\frac{\mathbf{Q}_1\mathbf{K}_1^\top}{\sqrt{d}}\right) - \lambda_1 \odot \text{Softmax}\left(\frac{\mathbf{Q}_2\mathbf{K}_2^\top}{\sqrt{d}}\right), \tag{6}$$

where $\lambda_1$ is a learnable parameter controlling the noise cancellation strength.

This subtraction filters noise inherent to the available modality. Specifically, guided by the overall survival objective, the linear projections are adaptively optimized during end-to-end training to capture these noisy representations. By canceling components consistently retrieved by both heads—specifically the dominant, uninformative background—it isolates pure, risk-relevant cues for downstream prediction. The refined differential feature $\mathbf{Z}$ is then computed as:

$$\mathbf{Z} = \text{Linear}(\mathbf{A} \cdot \mathbf{V}). \tag{7}$$

**Inference Strategy** The CDA module dynamically adapts based on data availability. In the **Complete Modality setting (Self-Refinement)**, features query their own prototype bank (e.g., $\mathbf{Q} = Project(\mathbf{F_{pa}}), \mathbf{K}, \mathbf{V} = Project(\mathcal{P}^{\mathbf{pa}})$) to sharpen risk patterns by subtracting background noise. In the **Missing Modality setting (Cross-Imputation)**, the available modality queries the missing modality's bank (e.g., $\mathbf{Q} = Project(\mathbf{F_{pa}}), \mathbf{K}, \mathbf{V} = Project(\mathcal{P}^{\mathbf{ge}})$). This retrieval-based strategy stably completes missing semantics by anchoring to reliable prototypes.

### 2.4 Semantic Rectification via Residuals

The refined feature $Z$ rectifies the available input $F$ via a residual connection ($F_{final} = F + Z$). In missing-modality scenarios, we directly output the imputed semantics ($F_{final} = Z$). To prevent training-inference distribution shifts, we apply random modality dropout during training, ensuring robust optimization solely on $Z$. This dynamically corrects noise and bridges modality gaps.

**Table 1.** Performance comparison using C-index (mean ± standard deviation) on five cancer datasets. The best result is shown in **bold**, and the second-best one is underlined. P and G are abbreviations of pathology and gene, respectively. $*$ represents the multimodal training and unimodal testing scenario (missing modality).

| Method | Modal | | Datasets | | | | | Overall |
|---|---|---|---|---|---|---|---|---|
| | P | G | BRCA | BLCA | STAD | COADREAD | HNSC | |
| ABMIL [7] | ✓ | × | $0.613_{\pm 0.118}$ | $0.575_{\pm 0.050}$ | $0.561_{\pm 0.059}$ | $0.586_{\pm 0.115}$ | $0.594_{\pm 0.036}$ | 0.586 |
| TransMIL [17] | ✓ | × | $0.625_{\pm 0.080}$ | $0.614_{\pm 0.046}$ | $0.522_{\pm 0.050}$ | $0.525_{\pm 0.118}$ | $0.573_{\pm 0.089}$ | 0.572 |
| WIKG [12] | ✓ | × | $0.613_{\pm 0.086}$ | $0.585_{\pm 0.047}$ | $0.551_{\pm 0.049}$ | $0.517_{\pm 0.127}$ | $0.581_{\pm 0.069}$ | 0.569 |
| MambaMIL [23] | ✓ | × | $0.617_{\pm 0.078}$ | $0.612_{\pm 0.058}$ | $0.574_{\pm 0.041}$ | $0.588_{\pm 0.095}$ | $0.531_{\pm 0.051}$ | 0.585 |
| SNN [10] | × | ✓ | $0.576_{\pm 0.112}$ | $0.539_{\pm 0.052}$ | $0.539_{\pm 0.032}$ | $0.616_{\pm 0.090}$ | $0.512_{\pm 0.076}$ | 0.556 |
| SNNTrans [10] | × | ✓ | $0.551_{\pm 0.097}$ | $0.553_{\pm 0.043}$ | $0.555_{\pm 0.057}$ | $0.614_{\pm 0.125}$ | $0.568_{\pm 0.031}$ | 0.568 |
| G-HANet [20] | ✓ | * | $0.677_{\pm 0.071}$ | $0.603_{\pm 0.054}$ | $0.588_{\pm 0.050}$ | $0.598_{\pm 0.169}$ | $0.602_{\pm 0.042}$ | 0.614 |
| ProSurv [14] | ✓ | * | $\underline{0.691}_{\pm 0.074}$ | $0.611_{\pm 0.032}$ | $0.620_{\pm 0.069}$ | $0.626_{\pm 0.090}$ | $0.576_{\pm 0.044}$ | 0.625 |
| Ours | ✓ | * | $0.663_{\pm 0.058}$ | $0.612_{\pm 0.030}$ | $\underline{0.626}_{\pm 0.021}$ | $\underline{0.686}_{\pm 0.078}$ | $0.587_{\pm 0.037}$ | 0.635 |
| ProSurv [14] | * | ✓ | $0.600_{\pm 0.150}$ | $0.643_{\pm 0.023}$ | $0.531_{\pm 0.106}$ | $0.618_{\pm 0.136}$ | $0.586_{\pm 0.044}$ | 0.596 |
| Ours | * | ✓ | $0.626_{\pm 0.089}$ | $\underline{0.678}_{\pm 0.067}$ | $0.604_{\pm 0.117}$ | $0.647_{\pm 0.118}$ | $\mathbf{0.652}_{\pm 0.024}$ | 0.641 |
| MCAT [3] | ✓ | ✓ | $0.654_{\pm 0.077}$ | $0.616_{\pm 0.055}$ | $0.572_{\pm 0.098}$ | $0.615_{\pm 0.141}$ | $0.560_{\pm 0.089}$ | 0.603 |
| MOTCat [21] | ✓ | ✓ | $0.665_{\pm 0.111}$ | $0.600_{\pm 0.051}$ | $0.563_{\pm 0.084}$ | $0.593_{\pm 0.169}$ | $\underline{0.630}_{\pm 0.021}$ | 0.610 |
| CMTA [27] | ✓ | ✓ | $0.604_{\pm 0.045}$ | $0.642_{\pm 0.072}$ | $0.577_{\pm 0.098}$ | $0.586_{\pm 0.075}$ | $0.550_{\pm 0.096}$ | 0.592 |
| SurvPath [8] | ✓ | ✓ | $0.675_{\pm 0.069}$ | $0.584_{\pm 0.040}$ | $0.551_{\pm 0.040}$ | $0.585_{\pm 0.100}$ | $0.598_{\pm 0.052}$ | 0.599 |
| PIBD [26] | ✓ | ✓ | $0.682_{\pm 0.092}$ | $0.634_{\pm 0.037}$ | $0.609_{\pm 0.076}$ | $0.637_{\pm 0.138}$ | $0.614_{\pm 0.111}$ | 0.635 |
| ProSurv [14] | ✓ | ✓ | $0.707_{\pm 0.082}$ | $0.662_{\pm 0.031}$ | $0.625_{\pm 0.089}$ | $0.650_{\pm 0.067}$ | $0.601_{\pm 0.054}$ | $\underline{0.649}$ |
| Ours | ✓ | ✓ | $\mathbf{0.717}_{\pm 0.051}$ | $\mathbf{0.683}_{\pm 0.056}$ | $\mathbf{0.649}_{\pm 0.083}$ | $\mathbf{0.687}_{\pm 0.066}$ | $0.624_{\pm 0.026}$ | **0.672** |

### 2.5 Optimization Objective

To extract discriminative risk patterns and prevent mode collapse, we introduce a Patch-Level Constraint:

$$\mathcal{L}_{patch} = \mathcal{L}_{align} + \lambda_2 \mathcal{L}_{orth} \tag{8}$$

where the alignment loss $\mathcal{L}_{align} = \sum_{x \in \mathcal{S}} \min_{p \in \mathcal{P}} \|x - p\|_2^2$ anchors prototypes $\mathcal{P}$ to high-risk regions by minimizing the distance between the bank and the set $\mathcal{S}$ of top-$k$ patches selected from uncensored patients. This exclusive focus on uncensored data ensures that the prototypes capture unambiguous prognostic signals, preventing the dilution of definitive high-risk features by the ambiguous, event-free patterns inherent to censored data. Additionally, the orthogonality loss

$\mathcal{L}_{orth} = \left\| \hat{\mathcal{P}}\hat{\mathcal{P}}^T - \mathbf{I} \right\|_F^2$ penalizes the normalized bank $\hat{\mathcal{P}}$ to maintain prototype diversity, and $\mathbf{I}$ is the identity matrix enforcing mutual orthogonality.

The overall objective function is:

$$\mathcal{L}_{total} = \mathcal{L}_{NLL} + \alpha\mathcal{L}_{sim} + \beta\mathcal{L}_{patch} \tag{9}$$

where $\alpha, \beta$ are hyperparameters, $\mathcal{L}_{NLL}$ is the standard negative log-likelihood [25], and where $\mathcal{L}_{sim} = \mathcal{L}_{sim}^{pa} + \mathcal{L}_{sim}^{ge}$ is the ordinal prototype alignment loss [14].

**Table 2.** Comprehensive ablation studies of ProtoDiffSurv. (a) Component-wise ablation under various modality settings (P: Pathology, G: Genomics; * indicates multimodal training with unimodal testing). (b) Comparison of patch weighting strategies ('One': uniform weight; 'Mean'/'Top': average similarity of all/top-5 prototypes; 'Max': maximum similarity).

| Method / Module | Modal | | Datasets | | | | | Overall |
|---|---|---|---|---|---|---|---|---|
| | P | G | BRCA | BLCA | STAD | COADREAD | HNSC | |
| **(a) Component Ablation** | | | | | | | | |
| w/o differential | ✓ | * | $0.651_{\pm 0.043}$ | $0.576_{\pm 0.043}$ | $0.567_{\pm 0.047}$ | $0.672_{\pm 0.108}$ | $0.540_{\pm 0.049}$ | 0.601 |
| w/o patch enhanced | ✓ | * | $0.657_{\pm 0.065}$ | $0.557_{\pm 0.045}$ | $0.557_{\pm 0.066}$ | $0.672_{\pm 0.084}$ | $0.511_{\pm 0.083}$ | 0.591 |
| Ours | ✓ | * | $0.663_{\pm 0.058}$ | $0.612_{\pm 0.030}$ | $0.591_{\pm 0.021}$ | $0.686_{\pm 0.078}$ | $0.537_{\pm 0.037}$ | 0.618 |
| w/o differential | * | ✓ | $0.599_{\pm 0.119}$ | $0.670_{\pm 0.069}$ | $0.563_{\pm 0.067}$ | $0.616_{\pm 0.113}$ | $0.616_{\pm 0.028}$ | 0.613 |
| w/o patch enhanced | * | ✓ | $0.604_{\pm 0.124}$ | $0.663_{\pm 0.072}$ | $0.615_{\pm 0.101}$ | $0.653_{\pm 0.114}$ | $0.630_{\pm 0.041}$ | 0.633 |
| Ours | * | ✓ | $0.626_{\pm 0.089}$ | $0.678_{\pm 0.067}$ | $0.604_{\pm 0.117}$ | $0.647_{\pm 0.118}$ | $0.652_{\pm 0.024}$ | 0.641 |
| w/o differential | ✓ | ✓ | $0.687_{\pm 0.065}$ | $0.669_{\pm 0.064}$ | $0.599_{\pm 0.052}$ | $0.651_{\pm 0.131}$ | $0.609_{\pm 0.016}$ | 0.643 |
| w/o patch enhanced | ✓ | ✓ | $0.693_{\pm 0.062}$ | $0.673_{\pm 0.058}$ | $0.626_{\pm 0.086}$ | $0.677_{\pm 0.073}$ | $0.622_{\pm 0.023}$ | 0.658 |
| Ours | ✓ | ✓ | $0.717_{\pm 0.051}$ | $0.683_{\pm 0.056}$ | $0.649_{\pm 0.083}$ | $0.687_{\pm 0.066}$ | $0.624_{\pm 0.026}$ | 0.672 |
| **(b) Patch Weighting Strategy** | | | | | | | | |
| one | ✓ | * | $0.657_{\pm 0.065}$ | $0.609_{\pm 0.024}$ | $0.557_{\pm 0.066}$ | $0.672_{\pm 0.084}$ | $0.491_{\pm 0.076}$ | 0.597 |
| mean | ✓ | * | $0.640_{\pm 0.072}$ | $0.602_{\pm 0.054}$ | $0.546_{\pm 0.117}$ | $0.660_{\pm 0.105}$ | $0.459_{\pm 0.079}$ | 0.581 |
| top | ✓ | * | $0.666_{\pm 0.074}$ | $0.596_{\pm 0.030}$ | $0.542_{\pm 0.087}$ | $0.666_{\pm 0.095}$ | $0.478_{\pm 0.060}$ | 0.590 |
| max | ✓ | * | $0.663_{\pm 0.058}$ | $0.612_{\pm 0.030}$ | $0.591_{\pm 0.021}$ | $0.686_{\pm 0.078}$ | $0.537_{\pm 0.037}$ | 0.618 |
| one | * | ✓ | $0.604_{\pm 0.124}$ | $0.665_{\pm 0.072}$ | $0.562_{\pm 0.149}$ | $0.653_{\pm 0.114}$ | $0.630_{\pm 0.041}$ | 0.623 |
| mean | * | ✓ | $0.566_{\pm 0.156}$ | $0.617_{\pm 0.092}$ | $0.551_{\pm 0.122}$ | $0.606_{\pm 0.133}$ | $0.613_{\pm 0.015}$ | 0.591 |
| top | * | ✓ | $0.600_{\pm 0.150}$ | $0.656_{\pm 0.092}$ | $0.566_{\pm 0.143}$ | $0.649_{\pm 0.146}$ | $0.619_{\pm 0.014}$ | 0.618 |
| max | * | ✓ | $0.626_{\pm 0.089}$ | $0.678_{\pm 0.067}$ | $0.604_{\pm 0.117}$ | $0.647_{\pm 0.118}$ | $0.652_{\pm 0.024}$ | 0.641 |
| one | ✓ | ✓ | $0.625_{\pm 0.111}$ | $0.673_{\pm 0.061}$ | $0.592_{\pm 0.146}$ | $0.677_{\pm 0.073}$ | $0.622_{\pm 0.023}$ | 0.638 |
| mean | ✓ | ✓ | $0.586_{\pm 0.144}$ | $0.656_{\pm 0.086}$ | $0.601_{\pm 0.116}$ | $0.675_{\pm 0.089}$ | $0.620_{\pm 0.014}$ | 0.628 |
| top | ✓ | ✓ | $0.683_{\pm 0.107}$ | $0.663_{\pm 0.087}$ | $0.623_{\pm 0.132}$ | $0.679_{\pm 0.110}$ | $0.622_{\pm 0.017}$ | 0.654 |
| max | ✓ | ✓ | $0.717_{\pm 0.051}$ | $0.683_{\pm 0.056}$ | $0.649_{\pm 0.083}$ | $0.687_{\pm 0.066}$ | $0.624_{\pm 0.026}$ | 0.672 |

## 3 Experiments and Results

We evaluated on five TCGA cohorts: BRCA (868), BLCA (359), STAD (318), HNSC (392), and COADREAD (294), reporting the average C-index from 5-fold cross-validation (6:2:2 split). Our ProtoDiffSurv, implemented in PyTorch with Swin Transformer features, was trained on an RTX 3090 GPU for 30 epochs

via Adam (lr=1e-5), with 4,096 patches sampled per slide following standard practice [3, 14].

For baselines, we adopt official implementations with their recommended settings. ABMIL, TransMIL, MCAT, MOTCat, CMTA, and PIBD use ImageNet-pretrained ResNet-50; WIKG and G-HANet use ResNet-50 with graph aggregators; MambaMIL, SurvPath, and ProSurv use CTransPath. Genomic data are processed as pathway-level RNA-Seq profiles. Unimodal genomic baselines (SNN, SNNTrans) use 3-layer MLPs. MCAT, MOTCat, PIBD, and ProSurv use Adam (lr=1e-4); CMTA uses SGD (1e-3); SurvPath uses AdamW (2e-4). Missing-modality evaluation follows the protocol of [3, 14].

### 3.1 Results

**Performance Comparison.** As shown in Table 1, ProtoDiffSurv achieves competitive or superior performance across diverse settings. In the Multimodal Setting, it dominates 4/5 datasets (overall 0.672), outperforming ProSurv (0.649) and PIBD (0.635) by effectively filtering conflicting noise. It also excels in Missing-Modality scenarios—notably achieving 0.652 vs. 0.586 (ProSurv) on HNSC (only Genomics)—confirming that our Cross-Modal Differential Attention successfully imputes missing risk semantics from the prototype bank without raw data.

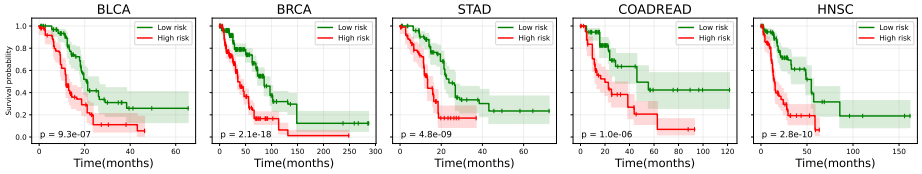


**Fig. 3.** Kaplan-Meier survival curves for five TCGA cancer cohorts stratified by predicted risk scores. Patients are divided into high-risk (red) and low-risk (blue) groups. All log-rank tests show statistically significant separation ($p < 0.001$).

**Ablation Studies.** Table 2a validates our dual-stage refinement: omitting Patch Enhancement degrades Pathology-only performance (confirming its spatial filtering role), while removing Differential Attention impairs Genomics-only and Multimodal results (proving its ability to cancel common noise). Furthermore, for patch weighting (Table 2b), explicitly selecting the most prototype-aligned patches ('Max') yields the best C-index, filtering background noise more effectively than averaging ('Mean') in sparse WSIs.

**Risk Analysis.** Kaplan-Meier curves [9] (Fig. 3) demonstrate that stratifying patients by predicted risk yields statistically significant separation ($p < 0.001$) across all five cohorts, confirming clinically meaningful prognosis.

While our missing-modality experiments follow the standard random dropout protocol [3, 14], real-world clinical missingness is often non-random, arising from

sequencing failures, insufficient tissue, or institutional heterogeneity. Such mechanisms may correlate with disease severity and introduce selection bias not fully captured by our simulation. Nevertheless, ProtoDiffSurv handles arbitrary missing patterns at inference without retraining, as each prototype bank is trained independently—a key advantage over generative imputation [16].

## 4 Conclusion

We propose ProtoDiffSurv, a dual-stage refinement framework for weakly supervised multimodal survival prediction. To handle the sparse risk signals in WSIs, we introduce Prototype-Guided Patch Enhancement. This module denoises patch features and highlights sparse regions that are most informative for risk. We further design a Cross-Modal Differential Attention mechanism to remove common noise and retrieve clean prognostic semantics from a prototype bank, and use them to refine the current features. Extensive experiments on five TCGA cohorts demonstrate ProtoDiffSurv's state-of-the-art performance and precise risk stratification in both complete and missing-modality settings.

**Acknowledgments.** The authors received no specific funding for this work.

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

## References

1. Cai, L., Huang, S., Zhang, Y., et al.: AttriMIL: Revisiting attention-based multiple instance learning for whole-slide pathological image classification from a perspective of instance attributes. Med. Image Anal. **103**, 103631 (2025)
2. Chen, C., Dou, Q., Jin, Y., et al.: Learning With Privileged Multimodal Knowledge for Unimodal Segmentation. IEEE TMI **41**, 621–632 (2022)
3. Chen, R.J., Lu, M.Y., Weng, W.H., et al.: Multimodal Co-Attention Transformer for Survival Prediction in Gigapixel Whole Slide Images. In: ICCV. pp. 4015–4025 (2021)
4. Clark, T.G., Bradburn, M.J., Love, S.B., et al.: Survival analysis part I: basic concepts and first analyses. Br. J. Cancer **89**, 232–238 (2003)
5. Dong, J., Jiang, J., Jiang, K., et al.: Disentangled Pseudo-Bag Augmentation for Whole Slide Image Multiple Instance Learning. IEEE TMI **44**, 4181–4197 (2025)
6. Hu, M., Maillard, M., Zhang, Y., et al.: Knowledge Distillation from Multi-modal to Mono-modal Segmentation Networks. In: MICCAI. pp. 772–781 (2020)
7. Ilse, M., Tomczak, J., Welling, M.: Attention-based Deep Multiple Instance Learning. In: ICML. pp. 2127–2136 (2018)
8. Jaume, G., Vaidya, A., Chen, R., et al.: Modeling Dense Multimodal Interactions Between Biological Pathways and Histology for Survival Prediction. In: CVPR (2024)
9. Kaplan, E.L., Meier, P.: Nonparametric Estimation from Incomplete Observations. J. Am. Stat. Assoc. **53**, 457–481 (1958)

10. Klambauer, G., Unterthiner, T., Mayr, A., et al.: Self-Normalizing Neural Networks. In: NeurIPS. pp. 1–10 (2017)
11. Krichen, M.: Generative Adversarial Networks. In: ICCCNT. pp. 1–7 (2023)
12. Li, J., Chen, Y., Chu, H., et al.: Dynamic Graph Representation with Knowledge-aware Attention for Histopathology Whole Slide Image Analysis. In: CVPR. pp. 11323–11332 (2024)
13. Li, R., Wu, X., Li, A., et al.: HFBSurv: hierarchical multimodal fusion with factorized bilinear models for cancer survival prediction. Bioinformatics **38**, 2587–2594 (2022)
14. Liu, F., Cai, L., Wang, Z., et al.: Prototype-Guided Cross-Modal Knowledge Enhancement for Adaptive Survival Prediction. In: MICCAI. pp. 522–532 (2025)
15. Lu, M.Y., Williamson, D.F.K., Chen, T.Y., et al.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nat. Biomed. Eng. **5**, 555–570 (2021)
16. Ning, Z., Zhao, Z., Feng, Q., et al.: Mutual-Assistance Learning for Standalone Mono-Modality Survival Analysis of Human Cancers. IEEE TPAMI **45**, 7577–7594 (2023)
17. Shao, Z., Bian, H., Chen, Y., et al.: TransMIL: Transformer based Correlated Multiple Instance Learning for Whole Slide Image Classification. In: NeurIPS. pp. 2136–2147 (2021)
18. Wang, Z., Gao, Q., Yi, X., et al.: Surformer: An interpretable pattern-perceptive survival transformer for cancer survival prediction from histopathology whole slide images. Comput. Methods Programs Biomed. **241**, 107733 (2023)
19. Wang, Z., Ma, J., Gao, Q., et al.: Dual-stream multi-dependency graph neural network enables precise cancer survival analysis. Med. Image Anal. **97**, 103252 (2024)
20. Wang, Z., Zhang, Y., Xu, Y., et al.: Histo-Genomic Knowledge Association for Cancer Prognosis From Histopathology Whole Slide Images. IEEE TMI **44**, 2170–2181 (2025)
21. Xu, Y., Chen, H.: Multimodal Optimal Transport-based Co-Attention Transformer with Global Structure Consistency for Survival Prediction. In: ICCV. pp. 21241–21251 (2023)
22. Xu, Y., Zhou, F., Zhao, C., et al.: Distilled Prompt Learning for Incomplete Multimodal Survival Prediction. In: CVPR. pp. 5102–5111 (2025)
23. Yang, S., Wang, Y., Chen, H.: MambaMIL: Enhancing Long Sequence Modeling with Sequence Reordering in Computational Pathology. In: MICCAI. pp. 296–306 (2024)
24. Ye, T., Dong, L., Xia, Y., et al.: Differential Transformer. In: ICLR (2025)
25. Zadeh, S.G., Schmid, M.: Bias in Cross-Entropy-Based Training of Deep Survival Networks. IEEE TPAMI **43**, 3126–3137 (2021)
26. Zhang, Y., Xu, Y., Chen, J., et al.: Prototypical Information Bottlenecking and Disentangling for Multimodal Cancer Survival Prediction. In: ICLR (2024)
27. Zhou, F., Chen, H.: Cross-Modal Translation and Alignment for Survival Analysis. In: ICCV. pp. 21485–21494 (2023)